

A FRAMEWORK FOR EXAMINING VALIDITY IN STATE ACCOUNTABILITY SYSTEMS

A PAPER IN THE SERIES:

IMPLEMENTING THE STATE ACCOUNTABILITY SYSTEM
REQUIREMENTS UNDER THE
NO CHILD LEFT BEHIND ACT OF 2001



February 2004



The Council of Chief State School Officers (CCSSO) is a bipartisan, nationwide, nonprofit organization of public officials who head departments of elementary and secondary education in the states, the District of Columbia, the Department of Defense Education Activity, and five U.S. extra-state jurisdictions. CCSSO provides leadership, advocacy, and technical assistance on major educational issues. The Council seeks member consensus on major educational issues and expresses their views to civic and professional organizations, federal agencies, Congress, and the public.

DIVISION OF STATE SERVICES AND TECHNICAL ASSISTANCE

The Division of State Services and Technical Assistance supports state education agencies in developing standards-based systems that enable all children to succeed. Initiatives of the division support improved methods for collecting, analyzing and using information for decision-making; development of assessment resources; creation of high-quality professional preparation and development programs; emphasis on instruction suited for diverse learners; and the removal of barriers to academic success. The division combines existing activities in the former Resource Center on Educational Equity, State Education Assessment Center, and State Leadership Center.

STATE COLLABORATIVE ON ASSESSMENT AND STUDENT STANDARDS

The State Collaborative on Assessment and Student Standards (SCASS) Program was created in 1991 to encourage and assist states in working collaboratively on assessment design and development for a variety of topics and subject areas. The Division of State Services and Technical Assistance of the Council of Chief State School Officers is the organizer, facilitator, and administrator of the projects.

SCASS projects accomplish a wide variety of tasks identified by each of the groups including examining the needs and issues surrounding the area(s) of focus, determining the products and goals of the project, developing assessment materials and professional development materials on assessment, summarizing current research, analyzing best practice, examining technical issues, and/or providing guidance on federal legislation. A total of forty-three states and three extra-state jurisdictions participated in one or more of the eleven projects offered during the project year 2003-2004.

COUNCIL OF CHIEF STATE SCHOOL OFFICERS

Ted Stilwill (Iowa), President

David P. Driscoll (Massachusetts), President-Elect

Michael E. Ward (North Carolina), Vice President

G. Thomas Houlihan, Executive Director

Julia Lara, Deputy Executive Director,
Division of State Services and Technical Assistance

Rolf Blank, Director of Education Indicators Programs and Coordinator
Accountability Systems and Reporting (ASR) SCASS

Jan Sheinker, Coordinator
Comprehensive Assessments Systems for ESEA Title I (CAS) SCASS

COUNCIL OF CHIEF STATE SCHOOL OFFICERS
ONE MASSACHUSETTS AVENUE, NW, SUITE 700
WASHINGTON, DC 20001-1431

(202) 336-7000
FAX (202) 408-8072
www.ccsso.org

Call (202) 336-7016 for additional information on CCSSO publications

A FRAMEWORK FOR EXAMINING VALIDITY IN STATE ACCOUNTABILITY SYSTEMS

A PAPER IN THE SERIES:
IMPLEMENTING THE STATE ACCOUNTABILITY SYSTEM REQUIREMENTS
UNDER THE NO CHILD LEFT BEHIND ACT OF 2001

Ellen Forte Fast and Steve Hebbler

with

ASR-CAS Joint Study Group on Validity in Accountability Systems

February 2004

ACCOUNTABILITY SYSTEMS AND REPORTING
COMPREHENSIVE ASSESSMENT SYSTEMS FOR ESEA TITLE I

State Collaboratives on Assessment and Student Standards

COUNCIL OF CHIEF STATE SCHOOL OFFICERS—WASHINGTON, DC

ISBN 1-884037-87-9

Financial support for the development of this paper came from the member states of the Accountability Systems and Reporting and the Comprehensive Assessment Systems for ESEA Title I State Collaboratives on Assessment and Student Standards (SCASS) projects. The Council of Chief State School Officers claims Copyright © 2004, for this material for the benefit of those member states.

Acknowledgements

This paper resulted from the work of the Joint Study Group on Adequate Yearly Progress (AYP) comprised of state education specialists and consultants from two SCASS projects: Accountability Systems and Reporting (ASR) and Comprehensive Assessment Systems for ESEA Title I (CAS). The members of the Study Group benefited tremendously from discussions among SCASS colleagues throughout 2003:

Reginald Allen, Minnesota
Jan Barth, West Virginia (co-Chair)
Wes Bruce, Indiana
Ron Carriveau, Arizona
H. Gary Cook, Harcourt
Tom Deeter, Iowa
Dorothy DeMars, Alabama
Steve Hebbler, Mississippi (co-Chair)
Ellen Hedlund, Rhode Island
Pat Roschewski, Nebraska
Ron Houston, Delaware
Ted Jarrell, Delaware

Robin Jarvis, Louisiana
Susan Ketchum, Wisconsin
Sandra McQuain, West Virginia
Les Morse, Alaska
Jason Nicholas, Wyoming
Kenna Seal, West Virginia
Alan Sheinker, CTB
Gary Skoglund, South Dakota
Christine Steele, Wyoming
Michael Taylor, Utah
Robin Taylor, Delaware
Charles Wayne, Pennsylvania
Jeffrey Zaring, Indiana

Rolf Blank, ASR SCASS Coordinator

Jan Sheinker, CAS SCASS Coordinator

Several others served as critical resources during the development of this paper. These include:

Bill Erpenbach, WJE Enterprises

Dale Carlson, StandBACC Consulting

J. P. Beaudoin, Research in Action

Paul LaMarca, Nevada Department of Education

Scott Marion, Center for Assessment

This publication and any comments, observations, recommendations, or conclusions contained herein reflect the work of the authors. They do not necessarily reflect the views of the Council of Chief State School Officers, its members, or the U.S. Department of Education.

Table of Contents

ACKNOWLEDGEMENTS.....	ii
TABLE OF CONTENTS	iii
LIST OF TABLES.....	v
LIST OF FIGURES.....	vi
PART I: FOUNDATIONS.....	1
OVERVIEW.....	2
Foundations for the Paper.....	2
Organization of this Paper.....	3
WHAT ARE ACCOUNTABILITY SYSTEMS?.....	4
WHY VALIDATE ACCOUNTABILITY SYSTEMS?	6
A BRIEF CONSIDERATION OF VALIDITY AND VALIDATION.....	9
Validity and Validation as Related to Assessments	11
EVOLUTION OF THE UNIFIED THEORY	12
FROM TESTS TO TESTING	14
CONSIDERATION OF CONSEQUENCES.....	14
VALIDATION AS EVALUATION	15
THE BASIC ASSESSMENT VALIDITY PREMISES	15
Assessment Validity Definition	16
Assessment Validity Caveats.....	16
Four Major Events in the Shifting Assessment Validity Paradigm.....	16
The Concept of Validity as it Relates to Accountability Systems.....	16
THE BASIC ACCOUNTABILITY SYSTEM VALIDITY PREMISES.....	17
Accountability System Validity Definition.....	17
Accountability System Validity Caveats	17
Translation for Accountability of the Four Major Events	17
PART II: THE VALIDATION FRAMEWORK	18
OUTLINE OF THE VALIDATION FRAMEWORK	18
The Basic Elements.....	18
Notes on Design and Data Collection	22
SECTION 1	
MAPPING THE SYSTEM	23
Clarifying Goals and the Theory of Action	23
Purpose.....	23
Major questions.....	23
SECTION 2	
EVALUATING THE INDICATORS.....	27
Purpose.....	27
Major questions.....	27
Indicators in Accountability Systems.....	27
Evaluating Test-Based Indicators	30
CONTENT AND CONSTRUCT QUESTIONS	32
Alignment.....	32
Response Processes.....	34
Comparability.....	34
RELIABILITY.....	35
Evaluating Rate Indicators	36

ALIGNMENT OF PRACTICE WITH DEFINITIONS	36
RELIABILITY	39
SECTION 3	
EVALUATING THE DECISION RULES.....	41
Purpose	41
Major questions	41
Introduction	41
Evaluating the Identification Process	41
DETERMINING THE LEVEL OF NEED FOR A SCHOOL	43
Using the Measures of Need	45
VALIDITY OF THE AYP MODEL.....	47
MEASURES OF NEED AND THE VALIDITY OF AYP MODELS.....	48
Stability and Reliability in the AYP Model.....	52
Misclassification Error and Validity of an AYP Model.....	54
EXAMPLES USING ACTUAL AYP RESULTS.....	57
SUMMARY AND CONCLUSIONS	62
Review and Appeals Processes.....	62
REVIEWS AND CORRECTIONS OF RAW DATA AND INDICATORS.....	63
REVIEWS AND APPEALS OF PRELIMINARY AYP RESULTS	63
SECTION 4	
EVALUATING THE CONSEQUENCES	64
Purpose	64
Major questions	65
Consequences in Accountability Systems	65
Examples of Consequential Validation Studies	68
CASE A: WERE THE SANCTIONS ACTUALLY IMPOSED AND WHO EXERCISES THE OPTION OF SCHOOL CHOICE?	68
Background.....	68
Design	68
CASE B: HOW IS CURRICULAR ALIGNMENT RELATED TO PERFORMANCE AND IMPROVEMENT STATUS?	71
Background.....	71
Design	71
CASE C: HOW IS TEACHER QUALITY AND THE QUALITY OF PROFESSIONAL DEVELOPMENT RELATED TO ACCOUNTABILITY PERFORMANCE?	74
Background.....	74
Design	74
Conclusions about the Study of Consequences.....	76
SUMMARY	76
BACKGROUND.....	76
FOUNDATIONS	77
Why Validate Accountability Systems?.....	77
What is Validity?	77
A VALIDATION FRAMEWORK	78
Clarifying Goals and the Theory of Action	78
Evaluating the Indicators	79
Evaluating the Decision Rules	79
Evaluating the Consequences	81
REFERENCES.....	82
APPENDIX A: CCSSO RESOURCES RELATED TO ACCOUNTABILITY SYSTEM VALIDATION	87
Resources on Alignment:	87
Resources on Validation of Accountability Systems	87
APPENDIX B: PROPOSED STANDARDS FOR EDUCATIONAL ACCOUNTABILITY SYSTEMS.....	88

List of Tables

TABLE 1.	COMPONENTS OF ACCOUNTABILITY SYSTEMS UNDER NCLB.....	4
TABLE 2.	EXAMPLE 5-YEAR PLAN FOR VALIDATION	21
TABLE 3.	GRAND THEORY OF ACTION QUESTIONS BY COMPONENT	24
TABLE 4.	WORKSHEET FOR INITIAL MAPPING OF MAJOR COMPONENTS AND THE THEORY OF ACTION	25
TABLE 5.	NCLB REQUIRED INDICATORS FOR AYP.....	28
TABLE 6.	OPTIONS FOR USING STATE-SELECTED INDICATORS IN ACCOUNTABILITY SYSTEMS	29
TABLE 7.	SAMPLE LISTING OF INDICATORS USED IN ACCOUNTABILITY DECISIONS	31
TABLE 8.	VALIDATION QUESTIONS AND STRATEGIES RELATED TO DATA	37
TABLE 9.	EXAMPLES OF VALIDITY-THREATENING DATA PROBLEMS AND THEIR RESOLUTIONS.....	38
TABLE 10.	EXAMPLE OF DIFFERENCE IN MAGNITUDE OF AMO MISSES.....	44
TABLE 11.	COMPARISON OF AVERAGE AND WEIGHTED AVERAGE DIFFERENCES IN GROUP AMO MISSES	45
TABLE 12.	COMPARISON OF THE AYP AND AAG MODELS UNDER PARADIGM 1.....	58
TABLE 13.	COMPARISON OF THE AYP AND AAG MODELS UNDER PARADIGM 2.....	59
TABLE 14.	COMPARISON OF AGREEMENT UNDER PARADIGMS 1 AND 2.....	59
TABLE 15.	COMPARISON OF THE AYP AND AAG MODELS USING THE THREE-CATEGORY SCENARIO.....	61
TABLE 16.	COMPARISON OF THE AYP AND AAG MODELS USING THE FIVE-CATEGORY SCENARIO	61
TABLE 17.	SAMPLE LISTING OF IMPOSED, EMERGENT, AND PLAUSIBLE NEGATIVE CONSEQUENCES.....	67
TABLE 18.	SAMPLE QUESTIONS AND EVIDENCE FOR CASE A	69
TABLE 18.	SAMPLE QUESTIONS AND EVIDENCE FOR CASE A (CONTINUED).....	70
TABLE 19.	SAMPLE STUDY PLAN FOR CASE B	73
TABLE 20.	SAMPLE STUDY PLAN FOR CASE C	75

List of Figures

FIGURE 1.	RELATIONSHIP BETWEEN COMPONENTS OF ACCOUNTABILITY SYSTEMS	6
FIGURE 2.	GRAPHIC REPRESENTATION OF CONSTRUCT UNDERREPRESENTATION AND CONSTRUCT IRRELEVANT VARIANCE	14
FIGURE 3.	GRAY AREA REGARDING ASSESSMENT VALIDATION PROCESS	15
FIGURE 4.	RELATION OF ACCOUNTABILITY SYSTEMS COMPONENTS VIA THE GRAND THEORY OF ACTION	19
FIGURE 5.	SAMPLE MODEL FOR MAPPING THE RELATIONSHIPS AMONG COMPONENTS	26
FIGURE 6.	SAMPLE FOUR-FOLD TRUTH TABLE FOR AYP	42
FIGURE 7.	MODEL FOR ANALYZING AGREEMENT BETWEEN TWO MODELS.....	51
FIGURE 8.	MODEL FOR ANALYZING AGREEMENT BETWEEN TWO YEARS	54
FIGURE 9.	EXAMPLE OF THREE- AND FIVE-CATEGORY GROUPINGS OF AYP RESULTS	60

Part I: Foundations

Edie Gonsalves has had a busy month. The final state test results came back from the contractor a few weeks ago after having been verified by each of the LEAs. The attendance rate files were completed about a week later; she'd had the graduation rates ready for a few months, since those are lagged by a year. Two weeks ago, with all of the pieces finally in hand, she ran the preliminary Adequate Yearly Progress (AYP) analyses. She's checked and rechecked the numbers and now sits with a few of her colleagues from the assessment and Title I units in the Superintendent's office, waiting to go over the final list of schools that have been identified for improvement.

They all wonder, "Are these the right schools?"

A scenario much like this one will play out at least once every year in states and local educational agencies (LEAs) across the country for the next several years. In most cases, the lists of schools identified for improvement will include schools that many would agree are troubled but also schools that most would agree are not. It is also likely that the set of schools not on the list will include both troubled and successful ones. Knowing this, state and LEA educators across the country will be asking themselves the same question asked by Edie and her colleagues: Are the right schools on the list?

Good question—and one that is not easy to answer. Whether the right schools were identified depends on what kinds of schools were meant to be identified. That, in turn, depends on the answer to the question at the very heart of the accountability system¹: “What are the goals that this accountability system is intended to achieve?”

In addition to this core question, Edie and her colleagues must answer several other questions before they can have any confidence in their list of schools or, just as importantly, evidence that their state's accountability system was working as they intended. For example, how trustworthy are the data on which the decision was based? How were these data combined in the decision-making model? And, what happens once schools are identified? What programs are implemented and how do we know they are appropriate and effective?

Clearly, the “right schools” question is only one small piece of a large and complex puzzle. To answer this question and the ones that precede, surround, and follow from it will require the accumulation of a large body of evidence and a thorough evaluation of that evidence. This paper focuses on how to build and evaluate that body of evidence: an evolving process that, when carried out systematically and rigorously, can be considered **validation**.

¹ The accountability systems addressed in this paper are standards-based and focus primarily on students' academic achievement rather than on compliance.

OVERVIEW

The purpose of this paper is to provide a framework for the evaluation of validity for accountability systems. Since neither validity nor accountability are particularly easy issues to tackle, attempting to create a framework for the consideration of one in the context of the other is risky business. Therefore, the framework offered here will almost certainly be incomplete and require revision over time. Likewise, the actual validation process for accountability systems will be tremendously arduous, although not nearly as problematic and costly as the alternative of doing nothing. Without validation evidence, states would be unable to effectively defend against lawsuits, would likely lose credibility among their stakeholders, and would almost certainly waste time and resources.

FOUNDATIONS FOR THE PAPER

1. **Every state must evaluate the validity of its accountability system.**

Either of its own volition or due to the requirements of the *No Child Left Behind Act of 2001* (NCLB), every state now operates an accountability system that imposes stakes—high stakes—on schools, school faculty, school LEAs, and perhaps also on students. Validation evidence is necessary to support the accountability claims made about individuals and agencies and the accompanying imposition of stakes. Professional standards for practice (APA/AERA/NCME, 1999; Baker, Linn, Herman, & Koretz, 2002) also highlight validation as being intrinsic to the decision-making process and the imposition of high stakes. The NCLB legislation itself makes 59 references to the need for validity with regard to assessment and/or accountability².

2. **States need accessible and flexible guidance on how to conduct this evaluation.**

Validity with regard to accountability systems has received little formal attention to date; at present, there exists no framework to guide states in carrying out this work.

This guidance must make technical validation and evaluation information accessible to state and local educators who may not have extensive training in measurement or research. Very few states employ a number of highly trained psychometricians and researchers whose sole job is to evaluate their assessments and accountability systems. Instead, most states employ a small number of professionals who are responsible for multiple programs and tasks. Though highly trained and experienced, these individuals generally do not have the time to sort out the basic premises for validation. It would also be a waste of resources for each state to approach this task separately. In December 2002, the Council of Chief State School Officers (CCSSO) published *Making Valid and Reliable Decisions in Determining Adequate Yearly Progress* (Marion, White, Carlson, Erpenbach, Rabinowitz, & Sheinker, 2002) as a first step in providing this guidance. The current paper takes the validity issues

² Principle 9 of the Consolidated State Application Accountability Workbook addressed issues of system reliability and validity. However, no state provided a plan for investigating these issues beyond the mention of data audits or appeals processes. To date, and with the exception of one state (Ohio), the US Department of Education (ED) has not requested any additional information about plans for examining reliability or validity (Erpenbach, Forte Fast, & Potts, 2003).

further and will, no doubt, be followed by subsequent papers that continue the conversation.

This validation guidance must also be flexible. States' accountability systems vary in a number of respects and guidance that cannot be tailored to fit each state's needs would be mostly useless. To ensure that the framework provided in this paper is flexible and widely applicable, representatives from many states were involved in every stage of its development and review.

3. **The validation process as described in this paper is basically grounded in the theories and methods for evaluating test validity.** This seems appropriate because they share a central premise—that a score serves as the basis for inferences and subsequent actions. However, assessments and accountability systems differ in ways that require the reinterpretation and extension of assessment validation practices for accountability systems. These differences will be described and addressed as they relate to the validation process.

ORGANIZATION OF THIS PAPER

This paper is organized into two parts.

Part I provides the foundation for the validation framework by considering, in turn—

- ♦ a definition of accountability systems,
- ♦ the reasons these systems require validation,
- ♦ a brief overview of validity and validation as represented in assessment, and
- ♦ the accountability system validation framework: application and extension of the assessment validation model for accountability systems.

Part II lays out in more detail each component of the validation framework.

Two appendices are included at the end of this paper:

A: CCSSO Resources Related to Accountability System Validation

B: Standards for Educational Accountability Systems

WHAT ARE ACCOUNTABILITY SYSTEMS?

Accountability systems can be defined in the following way:

Accountability systems are used to achieve specific educational goals by attaching to performance indicators certain consequences meant to effect change in specific areas of functioning.

Although that seems to be a simple statement, it actually comprises five critical concepts. First, the idea of a system, which, according to Merriam-Webster, is an “interdependent group of items forming a unified whole.” In this case, the “items that form the whole” are the other four critical concepts from the definition:

1. performance indicators
2. decisions rules
3. consequences
4. goals

To clarify what is meant by these concepts, consider them in the context of the *No Child Left Behind Act of 2001* (NCLB).

TABLE 1. COMPONENTS OF ACCOUNTABILITY SYSTEMS UNDER NCLB

Concept	Meaning under NCLB
Performance indicators	Indicators of – Percent Proficient in Reading or Language Arts Percent Proficient in Mathematics Participation Rate in Reading Participation Rate in Mathematics Graduation Rate Other Academic Indicators
Decision rules	AYP model: the decision rules that govern how scores are combined and then interpreted
Consequences	Imposed: • State defined rewards for high levels of performance • Sanctions for improvement schools, such as choice and reconstitution • Interventions to support school improvement Emergent (partial list; some only implicit in the law): • Changes in professional development systems • Changes in instructional methods • Changes in resource allocation
Goals	[For all students to attain reading and mathematics skills expected for their grade levels.] AMOs and IGs leading to 100% Proficiency in Reading/Language Arts by 2013-14 AMOs and IGs leading to 100% Proficiency in Mathematics by 2013-14

It is important to note that many people think of “consequences” as meaning something bad. That is not what is meant by this term here. Rather, consequences are simply one set of conditions that follow from another set of conditions. As Reckase (1998, p. 14) legitimately points out, consequences are more technically defined in a dictionary (Random House, 1987) as “the effect, result, or outcome of something occurring earlier,” strongly suggesting that consequences are the effect portion of a cause and effect relationship. As will be discussed briefly

below, limiting the meaning of consequences to “effects” can be problematic in the accountability context.

For the present purposes, it may be helpful to distinguish between two types of consequences within accountability systems, imposed and emergent. Imposed consequences, as defined here, are those that are imposed by a state or LEA according to results of the accountability decisions—they are consequences of performance as captured by the AYP model. These include rewards (e.g., cash incentives for faculty, release from procedural requirements), sanctions or “punishments” (e.g., loss of funds, provision of choice, reconstitution), and interventions or “help” (e.g., required partnerships with universities, master teachers, or mentor teams funded by state funds).

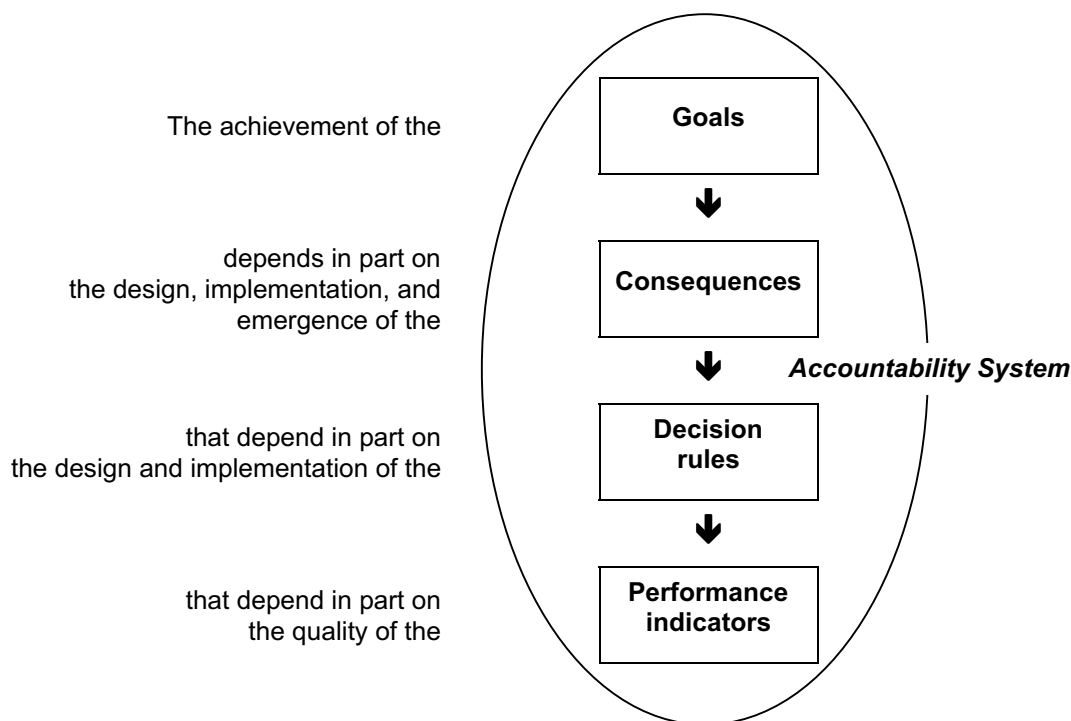
Emergent consequences are the activities and conditions in a school, LEA, or other entity that either occur in anticipation of imposed consequences or follow them chronologically. Changes in resource allocations or professional development practices are some possible emergent consequences. Negative, unintended consequences (e.g., use of inappropriate test preparation techniques, loss of experienced faculty) would also be considered emergent consequences. Emergent consequences are related to imposed consequences as specified in the Theory of Action.

Imposed consequences could be seen as effects: If a school does not meet its AYP target, then a specific sanction is imposed. But, this relationship might be considered more tautological than cause-and-effect in a pure research sense. Whether emergent consequences are effects (and what they are effects of) would be a matter for empirical study. However, it would not be possible to conduct the type of investigation necessary to test causality since schools cannot be randomly assigned to accountability conditions; doing so would be research but it would no longer be accountability. Thus, evaluations of these relationships will not meet the conditions required to establish causality.

The relationships between all four concepts (goals, consequences, decision rules, and performance indicators) could be represented as illustrated in Figure 1. When represented this way, it becomes clear that achievement of the end goals depends, in part, on the appropriate functioning of each of the other components. Further, unless each of these components is understood, it would be impossible to know why the goals were or were not achieved. Therefore, none of these components can be considered in isolation; the ultimate success of the system depends, in part, on the appropriate functioning of each. Evaluation of the system, then, must involve studies of each of the components and how they relate to one another.

In addition to these internal components, it is important to consider that accountability systems do not operate in a vacuum. These systems are situated within complex combinations of educational, political, and historical contexts. NCLB itself grew out of a specific national political-educational reform context and, as educators well know, must be implemented in a wide range of local political-educational circumstances in which student populations, resources, capacities, motivations, reform histories, and many other characteristics vary and interact. A framework for gaining a thorough understanding of these contexts is beyond the scope of this paper. However, serious consideration of these issues is encouraged and, where possible, suggestions are provided in this paper for how to do this.

FIGURE 1. RELATIONSHIP BETWEEN COMPONENTS OF ACCOUNTABILITY SYSTEMS



WHY VALIDATE ACCOUNTABILITY SYSTEMS?

Recalling the vignette that opened this paper, it is quite understandable that Edie and her colleagues may have some concerns about their identification of schools as needing improvement. They are concerned about identifying the “right schools” because this identification leads to the assignment of serious consequences. Under NCLB, these consequences include the distribution of rewards, the implementation of interventions, and the assignment of the following progressively severe sanctions, at least for schools that receive Title I funding:

- ♦ “A Title I school that has not made adequate yearly progress, as defined by the state, for two consecutive school years...will develop a two-year plan to turn around the school” and “students must be offered the option of transferring to another public school in the LEA—which may include a public charter school—that has not been identified as needing school improvement.”
- ♦ “If the school does not make adequate yearly progress for three years...the LEA must continue to offer public school choice to all students” and “students from low-income families are eligible to receive supplemental educational services, such as tutoring or remedial classes, from a state-approved provider.”
- ♦ “If the school fails to make adequate progress for four years, the LEA must implement certain *corrective actions* to improve the school, such as replacing certain staff or fully implementing a new curriculum, while continuing to offer public school choice and supplemental educational services for low-income students.”

- ♦ “If a school fails to make adequate yearly progress for a fifth year, the school/LEA must initiate plans for *restructuring* the school. This may include reopening the school as a charter school, replacing all or most of the school staff or turning over school operations either to the state or to a private company with a demonstrated record of effectiveness.”

Response by the U.S. Department of Education to the Frequently Asked Question, “What if a school does not improve?” Retrieved from <http://www.ed.gov/nclb/accountability/schools/accountability.html> (Emphasis in original)

Schools, faculty, and school/LEAs faced with these sanctions, as well as those who do not get rewards they believe they have earned, have a right to know how these decisions were made. The agencies making the decisions and imposing the consequences have a responsibility for communicating this information and for providing evidence to support their decisions (Baker & Linn, 2002; Baker, Linn, Herman, & Koretz, 2002). This evidence must extend far beyond the explanation of the algorithms used to score and categorize schools. It must encompass the goals of the system, the selection and production of the performance indicators, the decision rules of which the algorithms are a part, and the rationales for each of the imposed consequences.

To ground this statement in the big accountability picture, consider again the definition of accountability systems offered at the beginning of this section:

Accountability systems are used to achieve specific educational goals by attaching to performance indicators certain consequences meant to effect change in specific areas of functioning.

With that definition in mind, let’s say Edie and her colleagues work in a state that has implemented an accountability system with the purpose of achieving 100% proficiency in reading and mathematics among all students within 12 years.

What if, at the end of the 12th year, the goals are not achieved? What does this mean? Perhaps,

- ♦ the goals or the plan for reaching them were flawed?
 - Was the definition of proficiency too challenging?
 - Did the plan fail to address critical reform elements?
- ♦ the wrong schools were identified?
 - Were the “wrong” indicators used?
 - Were the indicators flawed in some way?
 - Were the indicators combined in a way that emphasized the wrong characteristics?
 - Was the decision model not sensitive to the right school characteristics?
 - Was the decision model too imprecise due to measurement or sampling error?
- ♦ the imposed consequences were ineffective?
 - Were imposed consequences, including rewards, sanctions, and interventions, assigned inappropriately?
 - Were the rewards not motivating or the sanctions and interventions ineffectively designed?

- Were imposed consequences poorly implemented?
- Were too many schools identified for the support systems to be effective?

These are some of the possibilities that may cause the system not to achieve the stated goals. But, unless Edie's state had been methodically studying its accountability system over the years, they would have no way to answer these questions. Indeed, if they had been studying their system, they might have recognized and addressed issues earlier in implementation. Alternatively, they might also have discovered what elements of their system were working; in fact, if their goals were achieved, only by studying their system would they understand why.

The crux of the problem is that it is impossible to explain or identify why a system does or does not work—or to defend the system against perfectly reasonable and inevitable challenges from affected stakeholders—unless it has been evaluated systematically.

To begin the conversation regarding how to conduct this systematic evaluation, consider, again, the definition of accountability systems from the previous page:

Accountability systems are used to achieve specific educational goals by attaching to performance indicators certain consequences meant to effect change in specific areas of functioning.

Assuming that the intended goals (e.g., the acquisition of adequate reading and math skills among all students) are both positive and valuable³, whether or not an accountability system works relies on three fundamental points:

1. the performance indicators are meaningful and relevant
and
2. the decision rules by which performance indicators are combined and attached to consequences function as intended
and
3. the imposition of specific consequences can ultimately lead to the intended goals by instigating or supporting intended reform activities; further, the imposition of consequences will not lead to unintended, serious, and negative changes and is preferable to doing either nothing or something else.

The crux of the problem is that it is impossible to explain or identify why a system does or does not work—or to defend the system against perfectly reasonable (and inevitable) challenges from affected stakeholders—unless it has been evaluated systematically.

Regardless of how obvious these points may seem, their truth as they relate to a given accountability system cannot be assumed. Rather, these points are tentative statements that must be tested empirically; they are hypotheses. In addition, these hypotheses are linked to each other and to the implemented elements of an accountability system via a theory of action: the

detailed, logical statement of how the accountability system is meant to work. Every agency that uses an accountability system to make judgments about

³ Whether specific goals are worth pursuing may be debatable, but such debate is not relevant to considerations of whether the system is working to achieve these goals (Shepard, 1993).

programs, groups, or individuals—and especially when those judgments lead to high stakes consequences such as those associated with most states’ accountability systems—should rigorously and systematically test these hypotheses and the theory of action in which they are embedded.

Why?

If the agency merely assumed that these hypotheses were true and the theory of action sound, it would lack evidence that it identified the “right” schools and that its decisions to reward or sanction students, schools, or LEAs were accurate and would lead to meaningful improvements in the intended (positive and valuable) educational outcomes. In other words, the agency would not be able to support the imposition of its consequences. Thus, the agency would not be able to defend its decisions in either public or legal forums and would almost certainly lose credibility with its stakeholders and challenges in court. Further, the agency could be wasting time and resources by investing in a system with substantial flaws and by not being able to identify and fix these flaws. If the system fails, it would be impossible to know why. It would also be impossible to know why a successful system worked so that the effective elements could be maintained and replicated.

Testing these hypotheses and the theory of action that subsumes them is a major endeavor. It cannot be done with a single study. Rather, it would require a comprehensive evaluation process. As an evaluation, the process would involve a series of studies, the collection and consideration of a wide range of evidence, and judgments about individual components of the system. In the context of educational assessment, this process would be characterized as validation⁴.

The next section encompasses a brief background on the concept of validity. This section on validity is followed by an overview of the approach to validation offered in this paper. This approach is then described in more detail in subsequent sections.

A BRIEF CONSIDERATION OF VALIDITY AND VALIDATION

*“Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the **adequacy and appropriateness of inferences and actions** based on test scores or other modes of assessment.”*

(Messick, 1989, p. 13)

emphasis in original

Those who impose accountability systems are the ones who must evaluate them; they are the ones setting the goals, making the claims, and distributing the consequences. For the present purposes, “they” are the staff of state departments of education and local education agencies.

That having been said, it would be inappropriate to proceed without mentioning the layers of other responsibilities within any education accountability system⁵. For example, the US Department of Education is responsible for monitoring the implementation of its policies and for providing funding and guidance, among other things. States are responsible for implementing both state and federal

⁴ The perspective that underlies the approach offered in this paper is that validation is evaluation (see Cronbach, 1988; Guion, 1980; Messick, 1980; and Shepard, 1993 for discussions of this perspective).

⁵ Personal communication with Dale Carlson (November, 2003).

policies—and for evaluating whether and how well these policies work in their local contexts. Though specific responsibilities may vary, local administrators are typically responsible for many aspects of accountability implementation, from assessment administration to data collection and management to the on-going development of their teachers' professional knowledge and skills. Teachers are responsible for developing a strong repertoire of instructional strategies and expertise in their content area(s). Students are responsible for showing up, working hard, and staying engaged. Parents are responsible for supporting their children in these endeavors.

The degree to which these responsibilities are borne by these parties will certainly affect the operation of the system—so responsibility for system functioning is diffused. But, at the end of the day, it is the state agency who determines and imposes policy requirements and so it is that agency that must evaluate them. This responsibility is not diminished because states' policies are constrained or extended due to federal mandates. Rather, since states are the conduit for federal policy implementation, they could consider it their duty to gather evidence related to whether and how well specific aspects of these federal policies actually work.

To support their validation work, state and local education agency staff should have a good understanding of the basic concept of validity. Although a thorough discussion of this topic is well beyond the scope and purpose of this paper (and would encompass hundreds of pages in itself), this section provides a brief overview meant to support the methodological strategies that are subsequently proposed. The following are among some excellent papers that explore validity in depth and may be good additional resources for the reader:

- ♦ Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Erlbaum.
- ♦ Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education and Macmillan.
- ♦ Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education*, 19, 405-450.

In addition, the major professional statement on validity in relation to assessment is presented in the *Standards*. This book, and its subsequent revisions, should be considered required reading for those engaged in educational assessment and/or accountability:

American Psychological Association, American Educational Research Association, & the National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

VALIDITY AND VALIDATION AS RELATED TO ASSESSMENTS

“Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests.”

(APA/AERA/NCME, 1999, p. 9)

If one were to select a sample of psychometricians from each of the last five to ten decades and gather them together in, say, a bar, it is quite likely that all would drink a toast to validity as the paramount concept in the field of testing. However, a *mêlée* would ensue if they were asked to define what validity *is*.

This scenario points to an important caveat about validity: it is not fixed either in its definition or as a property of a test, a test score, or even of an interpretation, inference, or use of a test score. It continues to evolve⁶ as a concept and cannot be captured conclusively with regard to any test—or for any accountability system, for that matter.

Perhaps, then, it is best not to think of validity itself as the prey. Rather, what one should seek is a body of evidence that has been gathered systematically and which can be judged in relation to specific claims (validity arguments).

Second, with regard to these validity arguments, the purpose of a validation process is not to prove claims true, worthy of making, or socially valuable—it is to “clarify for a relevant community what [a claim] means, and the limitations of each interpretation” (Cronbach, 1988, p. 3). In part, this clarification may involve the consideration of multiple “plausible rival hypotheses” (Campbell, 1957) regarding how and why a system functions as it does; some of these may be rejected and others supported based on the evidence at hand. Still, none can be judged right, as in absolutely accurate or just. For accountability systems, this means that engaging in a validation process cannot yield a stamp of approval for the system or for the goals it is meant to achieve.

Third, “validation is never finished” (Cronbach, 1988, p. 5). An agency’s responsibility for validation starts the moment someone decides to design and implement an accountability system and continues as long as the consequences of that system are applied. Study beyond that point may be important for a comprehensive understanding of the system and its effects, but is less obviously the responsibility of the agency.

With these caveats in mind, it is time to characterize the current perspectives on validity and validation. This is achieved by highlighting major events in the shifting validity paradigm⁷ with an emphasis on the issues that will affect the validation framework for accountability systems.

...the purpose of a validation process is not to prove claims true, worthy of making, or socially valuable—it is to “clarify for a relevant community what [a claim] means, and the limitations of each interpretation” (Cronbach, 1988, p. 3)

An agency’s responsibility for validation starts the moment someone decides to design and implement an accountability system and continues as long as the consequences of that system are applied.

⁶ Some (e.g., Angoff, 1988; Shepard, 1993) have characterized the changes in validity conceptions over the last century as “evolution,” which seems appropriate given that the concept has gradually grown in complexity over the last century via processes that could be called, variously, adaptation and mutation.

⁷ These major events are based in part on Angoff’s (1988) analysis of the evolution of the validity concept.

EVOLUTION OF THE UNIFIED THEORY

In the first few decades of the 20th century, the basic question posed in test validation was, does this test “measure what it purports to measure?” (Garrett, 1937, p. 324). Depending on the kind of test involved, psychometricians⁸ would have answered that question in one of two ways. In most cases, tests were used to either predict later performance or to assign people to categories. Wielding the relatively new tool of correlation, psychometricians validated these tests as being useful instruments for a given **predictive** or **concurrent** purpose by demonstrating that test scores were correlated with a **criterion**, such as job performance, success in school or a training program, or scores on other tests.

For tests used to measure knowledge of **content**, such as achievement or proficiency tests, validity would depend on how well the set of items on the test represented the domain the test was supposed to measure.

These three types of validity—predictive, concurrent, and content—were described in a joint publication of the American Psychological Association (APA), the American Educational Research Association (AERA), and the National Council on Measurement Used in Education (now, NCME) published as a supplement to *Psychological Bulletin* in 1954⁹.

A fourth type, **construct** validity, was also represented in the 1954 publication, although as “a weak sister to the [other types], suggesting it as a substitute when a real criterion was not available” (Shepard, 1993, p. 416).

The following year marked a milestone in validity’s evolution. The chairman of the committee that wrote the 1954 publication, Lee Cronbach, and another member of that committee, Paul Meehl, published what has become a classic paper in the field (Cronbach & Meehl, 1955). Drawing from the logical positivism of the physical sciences, Cronbach and Meehl proposed the “nomological net” as a theoretical framework connecting the test scores, behaviors, and content domains that were the focus in validation at that time to the latent psychological constructs that underlie them. This model required specification and empirical testing of the relationships between hypothesized *constructs* and observed behaviors within this nomological net. Interpretations of test scores were validated through the patterns of confirmations or disconfirmations of specific, theory-based hypotheses. The specific hypotheses related to a given test interpretation were to address both the patterns internal to the test (e.g., correlations among items or subscales) and the patterns between the test score and external variables (e.g., performance on other tests or behavior in related circumstances).

Although Cronbach himself subsequently stated that the full application of this sort of positivist model in the behavioral sciences was “pretentious” (Cronbach, 1989, p. 159), the nomological net redefined validation as a theory-driven enterprise requiring a systematic approach and, thus, remains a guiding force in validation work today.

⁸ According to the Merriam-Webster on-line dictionary, the earliest known reference for the term “psychometricians” occurred in 1939. So, researchers/statisticians/ psychologists from before that time may not have referred to themselves as psychometricians, but they are referred to as such in retrospect.

⁹ This article, “Technical recommendations for psychological tests and diagnostic techniques,” is now considered the first edition of the Standards for Educational and Psychological Testing.

Cronbach and Meehl's conceptualization of validity and validation marked a shift in the validity paradigm. Although the 1966 *Standards for Psychological Tests and Manuals* (APA, 1966) portrayed three types of validity¹⁰, this "trinitarian" (Guion, 1980) perspective soon gave way. Through the 1970s and 1980s, several prominent psychometricians (e.g., Cronbach, 1971, 1988; Guion, 1977, 1980; Messick, 1980, 1988, 1989; Tenopir, 1977) embraced a central focus on construct validity and theory-driven approaches to validation. Addressed in the 1985 Standards (APA/AERA/NCME, 1985) and laid out in detail in Messick's epic chapter (Messick, 1989), this unified perspective unites all other 'types' of validity within construct validity. Criterion-related and content-related types of evidence are critical to the interpretation and use of a test score but these are no longer considered separate types of validity¹¹. As described by Messick (1980), "construct validity is...the unifying concept of validity that integrates content and criterion considerations into a common framework for testing rational hypotheses about theoretically relevant hypotheses" (p. 1015).

Validation from the unified perspective involves the collection and evaluation of evidence related to the two major classes of threats to validity: construct

underrepresentation and construct irrelevant variance (APA/AERA/NCME, 1999; Messick, 1989). With the former, an assessment is defined too narrowly and does not address the full breadth or depth of the intended construct; it would be inappropriate to interpret scores from such an assessment as representing the intended construct. Therefore, it could be inappropriate to use these scores to make decisions related to that construct.

With (construct underrepresentation), an assessment is defined too narrowly and does not address the full breadth or depth of the intended construct; it would be inappropriate to interpret scores from such an assessment as representing the intended construct.

Valid score interpretations are threatened by construct irrelevant variance when an assessment measures constructs in addition to the one targeted. This can occur in a number of ways, such as when test items meant to measure mathematics computation skills draw too heavily upon a student's reading skills or when students' test anxiety or poor eyesight interferes with their ability to answer the item correctly. It could be inappropriate to use a score based on such items to make decisions related to the intended construct.

Valid score interpretations are threatened by construct irrelevant variance when an assessment measures constructs in addition to the one targeted.

These threats are always present; the purpose of validation is to judge whether their influence renders a given score interpretation and use untenable.

¹⁰ In the 1966 Standards, three types of validity were represented: predictive and concurrent validity were subsumed under criterion validity and joined by content and construct.

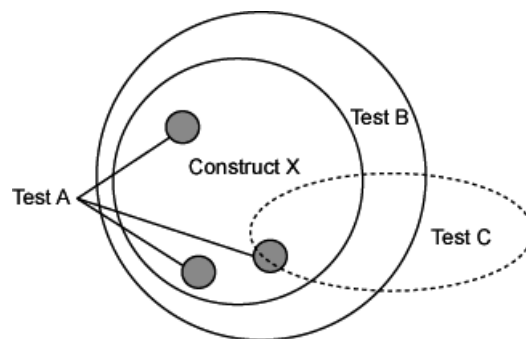
¹¹ Messick (1994) later distinguished a total of six aspects of construct validity: (1) content; (2) substantive; (3) structural; (4) generalizability; (5) external; and (6) consequential. Yet another "type" of validity, face validity, which is the perception on the part of the test taker that the test is measuring what it is supposed to be measuring, has received some attention over the years but was never raised to the same level of significance as the criterion, content, and construct "types."

FIGURE 2. GRAPHIC REPRESENTATION OF CONSTRUCT UNDERREPRESENTATION AND CONSTRUCT IRRELEVANT VARIANCE

Test A only reflects parts of Construct X, so it underrepresents this construct.

Test B measures constructs in addition to Construct X, so scores on this test will reflect construct irrelevant variance.

Test C both underrepresents Construct X and reflects construct irrelevant variance.



FROM TESTS TO TESTING

The 1985 *Standards* reflected not only the shift from the trinitarian to the unified perspective on validity, but also a shift in emphasis from instruments to practices.

This change was evidenced in the subtle yet important change in the document's

...validation now must address "the degree to which empirical evidence and theoretical rationales support the **adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment**" (Messick, 1989, p. 13).

title from the *Standards for Educational and Psychological Tests* to the *Standards for Educational and Psychological Testing* (Angoff, 1988; emphasis added). This shift in emphasis means that instead of merely addressing the question, "Does this test measure what it is purported to measure?", Or, as recast by Shepard (1993), "Does this test do what it claims to do?" rather than only "Does the test measure what it is supposed to measure?"

CONSIDERATION OF CONSEQUENCES

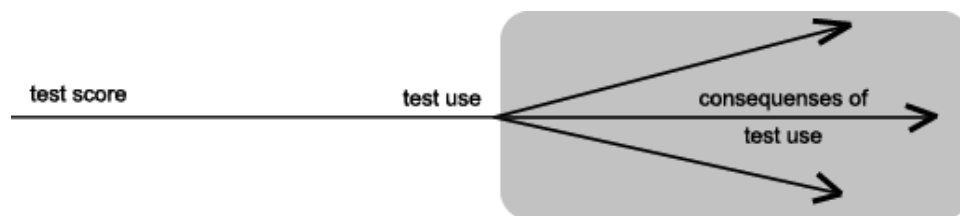
Following from the discussion thus far, it is clear that validity and validation relate not to a test score but to the use of that test score for a specific purpose (APA/AERA/NCME, 1999). Uses of test scores such as in making decisions about program eligibility or identifying schools for improvement have consequences such as receipt of services or school staff reconstitution. Does the web of validity extend to these consequences?

This is the subject of a current debate within the measurement community. At issue is whether the consequences of test use "should be an integral part of validity theory and practice" (Shepard, 1997, p. 5) or "needlessly complicate the conception of ...validity" (Wiley, 1991, p. 88). In both the major statement on the unified theory of validity and several subsequent papers, Messick (1989, 1992, 1994) argues that the social consequences of test score use must be evaluated to ensure that "adverse social consequences [are not] attributable to any source of test invalidity such as construct underrepresentation or construct irrelevant variance" (Messick, 1994, p. 24). The 1985 Standards indicate that consequences should be considered part of validation only if the test developer or the test user claims that the test can be used for engaging interventions. By 1999, the Standards also explicitly distinguished between evidence about those consequences that are "directly relevant to validity" and those "that may inform decisions about social policy but [fall] outside the realm of validity" (AERA/APA/NCME, 1999, p. 16).

The critical point for the present purposes is that it is conceptually possible to separate a test, the scores that result from its administration, and the use of those

scores from the consequences that result from the use of those scores. In other words, one could draw a line somewhere in the gray area below and consider everything to the left of that line an essential part of the assessment validation process.

FIGURE 3. GRAY AREA REGARDING ASSESSMENT VALIDATION PROCESS



Where one drew that line would depend on the test and a particular use of its scores, one's role in the testing process (e.g., test developer, test user), and one's position in this debate. But, that line would only represent what range of evidence should be considered as part of the validation process. There is no debate and no line to be drawn about whether the consequences associated with testing ought to be *evaluated*—just whether that evaluation is necessary for *validation*.

VALIDATION AS EVALUATION

In the days when criterion validity reigned supreme—alongside the somewhat more minor deity of content validity—the validation process would typically have involved the calculation of correlation coefficients. With the reconceptualization of validity as “the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (APA/AERA/NCME, 1999, p. 9), much more is required in a validation process.

Recognizing this, a number of prominent psychometricians (e.g., Cronbach, 1988, 1989; Guion, 1980; Messick, 1980; Shepard, 1993) have characterized validation as evaluation, which is the “systematic investigation of the worth or merit of an object” (Joint Committee on Standards for Educational Evaluation, 1999, p. 3). In the assessment case, the object would be a claim for using a particular test score in a particular way and “systematic” would refer to the methodical process applied to problem definition, study design, data collection, data analysis, and reporting.

Characterizing validation as evaluation has a number of benefits for validation practice. First, the principles and standards of program evaluation, which are comprehensive and methodologically sound, can guide the validation process. Second, thinking of validation as evaluation reaffirms that the overall goal of the process is not to establish proof but to gather evidence related to multiple strands that, together, represent a compelling argument regarding the program, or score, in question. Third, an evaluation framework provides practitioners with a structure for framing their questions. This allows them to connect the big picture to the details, prioritize which questions to address first, and to identify and address the weak links in their arguments.

THE BASIC ASSESSMENT VALIDITY PREMISES

The preceding section provided a definition, three caveats, and four major events in the shifting validity paradigm. These are summarized below:

Assessment Validity Definition

“Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (APA/AERA/NCME, 1999, p. 9).

Assessment Validity Caveats

- ♦ Validity is not a property of a test, a test score, or even of an interpretation, inference, or use of a test score. It cannot be captured conclusively. Rather, a judgment must be made regarding whether a body of evidence supports specific test claims and uses.
- ♦ A validation process cannot prove interpretations or uses of test scores true, worthy of making, or socially valuable.
- ♦ “Validation is never finished” (Cronbach, 1988, p. 5).

Four Major Events in the Shifting Assessment Validity Paradigm

- ♦ The evolution of the unified theory of validity, in which construct validity is the major concept;
- ♦ A shift in validation focus from testing instruments to testing practices, from “Does this test measure what it is purported to measure?” to “Does this test do what it claims to do?”; validation as addressing the potential threats to test score interpretations and uses;
- ♦ The debate regarding whether the consequences of test use should be considered as part of the validation process; and
- ♦ The characterization of validation as evaluation that draws upon the principles and practices of that field.

THE CONCEPT OF VALIDITY AS IT RELATES TO ACCOUNTABILITY SYSTEMS

“An accountability system can be said to have validity when the evidence is judged to be strong enough to support the inferences that:

[1] the components of the system are aligned to the purposes, and are working in harmony to help the system accomplish those purposes; and

[2] the system is accomplishing what was intended (and did not accomplish what was not intended.)”

(Marion et al., 2002, p. 105).

As suggested in the previous section, a substantial body of literature has accumulated over the years relating to assessment validity. Although the concept of validity is an evolving one and professionals in the field would never say that our understanding of validity or validation is (or could be) complete, it is fair to say that assessment validity and how to evaluate it has received a fair amount of attention.

This is far from true for accountability systems. As noted in a CCSSO publication that preceded the present one, *Making Valid and Reliable Decisions in Determining AYP* (Marion et al., 2002), “there is almost no literature on the

validity of accountability systems” (p. 38). The National Center for Research on Evaluation, Standards, and Student Testing (CRESST) has published a number of briefs and reports that relate to validity issues and standards for accountability systems (e.g., Baker & Linn, 2002; Baker, Linn, Herman, & Koretz, 2002). The National Center for the Improvement of Educational Assessment (NCIEA) also offers solid guidance and hands-on assistance for states with regard to these issues. These organizations will, no doubt, continue to serve as excellent resources for those who must implement and evaluate these systems. But, the fact remains that there is, as yet, no comprehensive model for validating accountability systems. So, using the principles and major issues described in the previous section, a foundation for such a model is offered here.

The validation framework for accountability systems that is offered here is grounded in the premises summarized in Basic Assessment in Validity Premises, above. However, since assessments—even assessment systems—are not the same thing as accountability systems, these premises must be extended and reshaped a bit to fit the problem at hand.

THE BASIC ACCOUNTABILITY SYSTEM VALIDITY PREMISES

The fundamental difference between assessments and accountability systems is that assessments are measurement tools employed for a specific purpose while accountability systems are systems that encompass measurement, evaluation, and reform that are employed to achieve certain goals. Standards-based educational accountability systems always subsume assessment systems.

With this in mind, as well as the definition of accountability systems presented earlier in this paper, the assessment validity premises can be recast for accountability as follows.

Accountability System Validity Definition

Validity with regard to accountability systems refers to the degree to which evidence and theory support the indicators, decisions, and consequences, individually and combined as established via the theory of action, as used for the purpose of achieving specific goals.

Accountability System Validity Caveats

- ♦ Validity is not a property of an accountability system or of a decision made as part of that system (e.g., this school needs improvement), and validity cannot be captured conclusively. Rather, a judgment must be made regarding whether a body of evidence supports the system and each of its components as implemented for the intended purpose.
- ♦ A validation process cannot prove accountability systems worthy of implementing, nor can the process prove any accountability decision true, worthy of making, or socially valuable.
- ♦ “Validation is never finished” (Cronbach, 1988, p. 5).

Translation for Accountability of the Four Major Events

- ♦ **The unified theory of validity**
For accountability systems, this means that all elements of the accountability system are interrelated according to a theory of action. No

element of the system can be considered in isolation, nor can evidence of the strength of any one aspect of the system cancel out weaknesses elsewhere—but a weakness in one component can certainly undermine a strength in another.

- ♦ **The shift in validation focus from testing instruments to testing practices, claims, and inferences**

Accountability is not Adequate Yearly Progress (AYP). The overarching validity question is not “Does this accountability system select the right schools?”, but rather, “Does this accountability system do what it is intended to do?” Selecting the right schools is only part of the answer.

- ♦ **The debate regarding consideration of consequences as part of test validation**

There is no debate about consideration of the consequences associated with accountability systems. Consequences that are imposed as part of the system—those that result from the imposition of rewards, sanctions, and interventions, and important unintended consequences—must all be considered as part of the validation process.

- ♦ **Validation as evaluation**

Accountability systems represent a claim that goal “X” can be achieved through the application of specific tools and activities. That claim must be evaluated to defend against competing claims.

Part II: The Validation Framework

The purpose of this part of the paper is to present the outline for an accountability system validation process framework. This outline is described in the first section below; each component is then described in more detail in subsequent sections. This framework is based on the foundations provided in Part I of this paper.

OUTLINE OF THE VALIDATION FRAMEWORK

The framework for validation presented here integrates the substance of accountability systems with processes of validation and evaluation, as generally established in the assessment and program evaluation contexts.

With regard to the substance of accountability systems, recall the definition of these systems from Part I:

Accountability systems are used to achieve specific educational goals by attaching to performance indicators certain consequences meant to effect change in specific areas of functioning.

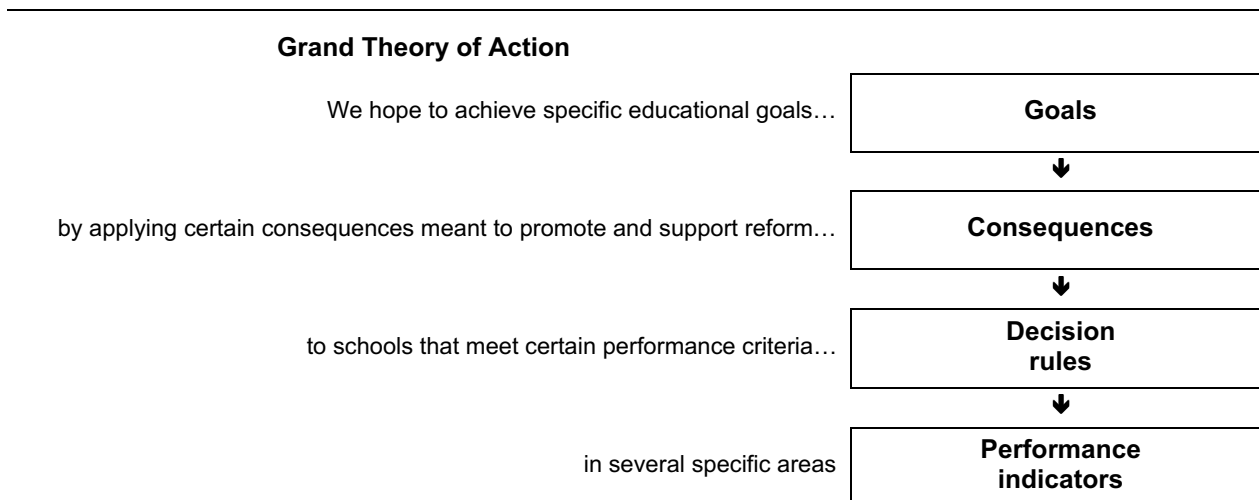
Process information follows. An important note: maintain clear records of the validation work you are doing and plan to do. That alone may help save credibility and even help to address some legal concerns that might arise in the future.

THE BASIC ELEMENTS

The basic elements of these systems, performance indicators, decision rules, consequences, and goals, are related to one another according to a grand **Theory**

of Action. This Theory of Action is the logic by which the system is intended to function; a simplified illustration follows.

FIGURE 4. RELATION OF ACCOUNTABILITY SYSTEMS COMPONENTS VIA THE GRAND THEORY OF ACTION



The validation framework presented here involves the evaluation of each of these elements as well as of the theory of action that underlies their association.

As for the validation process, the key premises are summarized at the end of Part I. The basic plan¹² is as follows:

1. Map the system

Identify the goals and specify the actions and logic by which the system is meant to achieve these goals.

2. Evaluate the indicators

Review the definitions and construction of each indicator both as intended and as implemented.

3. Evaluate the decision rules

Analyze the process by which indicators are combined and fed into decisions. Also, evaluate the decision outcomes in relation to other information about the schools and LEAs.

4. Evaluate the consequences

Analyze the full array of the consequences that are or could be applied to schools, evaluate how these consequences are actually implemented, and study how the application of these consequences plays out in both intended and unintended ways.

In practice, these pieces cannot be considered or conducted linearly; thus, they are not referred to here as “steps.” Ideally, system mapping would indeed be the first step in both the development and the evaluation of the system, but the speed with which states have had to design and implement accountability decisions, and

¹² It is important to remember that this is a plan for evaluating the system, not for developing it. See Gong (2002) for a thorough discussion of development considerations.

the prescriptive nature of NCLB requirements have pushed evaluation of the decision rules to the top of the evaluation priority list. That is, states must be evaluating their AYP decisions as soon as they make them. So, system mapping may need to wait in some cases.

In addition, some studies could be conducted concurrently. At the same time that states are making and evaluating their AYP decisions, they must be planning and collecting data for evaluations of the other accountability system elements (e.g., ensuring the quality of the indicators is critical to the functioning of the decision rules and the decision rules do not mean much if not attached to meaningful consequences). So, consideration of the indicators should not lag too far behind reviews of the AYP decisions. Also, the validation process never ends; completion of one study always leads into another study. Thus, design, data collection, analysis, and reporting functions that correspond to evaluations of different elements of the system will always occur simultaneously.

Given their limited resources, states may be best advised to prioritize their questions and to think of the “never-ending” validation process as a series of five-year plans.

Given their limited resources, states may be best advised to prioritize their questions and to think of the “never-ending” validation process as a series of five-year plans. Table 2 provides an example of how a state could begin to plot out its validation plan in a way that addresses each of the important issues over time. In some cases, issues are addressed through studies, which have a beginning and an end; in others,

validation means the development and implementation of processes and procedures that will continue indefinitely.

The example study plan presented in Table 2 illustrates a multi-method approach to the validation process. In combination with studies related to the assessment system, this sample includes two Advisory Panels, an annual survey, and a series of site visits. The Practitioner Advisory Panel, made up of teachers, parents, and administrators, would allow stakeholders an opportunity to have legitimate input into the accountability system and provide the state with useful information on implementation. Use of such a panel could also enhance the credibility of the state’s accountability plans. Much like the technical committees convened for assessment purposes, the Technical Advisory Panel would include experts who could review and advise on the states’ accountability system as well as the validation process itself.

TABLE 2. EXAMPLE 5-YEAR PLAN FOR VALIDATION

Issue	Year 1	Year 2	Year 3	Year 4	Year 5
Cross-issue activities	- Select an Advisory Panel of Practitioners (teachers, parents, and administrators) - Select an Advisory Panel of technical experts - Survey a sample of teachers and administrators	- Resample and resurvey - Convene advisory panels once each semester - Select site visit sample (coordinate design and visits across studies for other components)	- Resample and resurvey - Convene advisory panels once	- Resample and resurvey - Convene advisory panels once each semester	- Resample and resurvey - Convene advisory panels once
What are the goals of the system and how are they meant to be achieved?	Document system design process	- Include relevant questions on survey - Meet with advisory panels	Meet with advisory panels	- Include relevant questions on survey - Meet with advisory panels	Meet with advisory panels
How valid and reliable are the AYP decisions?	- Review the AYP calculations - Design and implement appeals procedures - Include relevant questions on survey	- Review the AYP calculations - Conduct site visits to gather information related to “right schools” question	- Review the AYP calculations - Include relevant questions on survey	- Review the AYP calculations - Conduct site visits to gather information related to “right schools” question	- Review the AYP calculations - Include relevant questions on survey
How valid and reliable are the AYP indicators?	- Revisit the data definitions and data collections related to AYP indicators - Work out assessment validation plan with contractor, including designation of data collection strategies and responsibilities - Conduct external alignment study of reading and mathematics assessments	- Conduct analyses related to assessment validity/reliability - Revisit achievement standards given new accountability uses - Conduct site visits to study data collection procedures and understanding of data definitions; use information from site visits to clarify definitions and procedures, highlighting common problems	- Conduct analyses related to assessment validity/reliability - Develop and disseminate revised data definitions and guidance on data collection procedures - Develop and implement auditing process for AYP indicators	- Conduct analyses related to assessment validity/reliability - Continue implementation of auditing - Monitor redevelopment of reading/language arts standards	- Conduct analyses related to assessment validity/reliability - Continue implementation of auditing - Monitor redevelopment of mathematics standards
How effective are the consequences in effecting the intended changes?	- Include relevant questions on survey - Design and implement monitoring process for schools in improvement	- Include relevant questions on survey - Conduct site visits to gather implementation data	- Include relevant questions on survey - Conduct site visits to gather implementation data - Conduct focus groups on reports and support systems	- Include relevant questions on survey - Conduct site visits to gather implementation data	- Include relevant questions on survey - Conduct site visits to gather implementation data - Conduct focus groups on reports and support systems

With regard to the survey, this sample plan would involve the selection of a sample of schools—and perhaps a sample of teachers within these schools—for survey administration. Depending on the questions the state hoped to address through the survey, the state could choose survey samples to represent a combination of low-performing and high-performing schools, a combination of grades, a range of urbanicity, or a number of other factors. One benefit of selecting a combination of low- and high-performing schools would be the ability to compare responses across these performance levels. This would be important if questions such as, “Do low- and high-performing schools differ with regard to participation in professional development activities?” However, even if the state were not primarily interested in such comparisons, it would be wise to select a sample that represented a range of schools rather than, for example, only low-performing schools in urban settings. A highly homogenous sample would make it difficult to interpret results as being related to anything other than the type of schools surveyed.

Similarly, a state would be wise to select a range of schools for the site visits. Only visiting the schools identified for improvement would limit the interpretability and usefulness of the results.

It is important for states to recognize that although validation requires the use of resources that could certainly be used elsewhere, *not* validating their systems could easily be far more expensive. As noted earlier in this paper, the possible waste of time and resources, the loss of credibility, and the risk of lawsuits without a proper defense would cost much more than validation work, especially when that work is efficiently designed.

NOTES ON DESIGN AND DATA COLLECTION

As noted above, some aspects of validation involve conducting studies that have a beginning and an end (e.g., an external study of the content alignment between the reading assessment and the reading standards). Other aspects involve the study and refinement of processes that are on-going and require on-going evaluation (e.g., reviews of the data collection or appeals protocols).

In either case, states should expect to use multiple data collection methods in their validation work and would be well advised to consult well-recognized resources when designing their studies and their data collection strategies. Some excellent resources include—

- ♦ Babbie E. R. (1998). *The basis of social research*, (8th ed.). Belmont, CA: Wadsworth.
- ♦ Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Boston: Houghton Mifflin.

Since most states will use site visits as part of their evaluation plans, a few comments regarding such visits seem warranted here:

- ♦ Site visits should be carefully planned so that the maximum amount of information can be collected in the least bothersome or intrusive way. Any site visit will disrupt the normal flow at the site—simply knowing that “the state” is coming will change some behaviors!—but the disruption to students, especially, should be minimized.
- ♦ If possible, two to three data collectors should visit a site at one time; each should have a particular focus and set of responsibilities. One may

interview the principal and three teachers. Another may review fiscal records while the third conducts two classroom observations. Such concurrent collection strategies will reduce the cost for the visit and also signal to those being visited that the state values their time.

- ♦ To further enhance efficiency, it may be possible to gather information related to several studies during a single site visit; information about data management and the implementation of consequences could be collected simultaneously. Or, evaluation site visits could be combined with visits for technical assistance purposes. However, activities related to each of these two purposes should be conducted separately during the visit. For example, in the evaluation portion of a visit, which might occur on Monday and Tuesday, staff could collect data. On Wednesday and Thursday, staff could engage in technical assistance activities. This way, personnel at the site would get some immediate feedback. State and regional staff who work with schools in their reform efforts may be excellent resources when designing, preparing for, and interpreting results from site visits.

SECTION 1

MAPPING THE SYSTEM: CLARIFYING GOALS AND THE THEORY OF ACTION

Purpose

To specify what the accountability system is meant to achieve and the means by which these goals are to be reached. Also, to identify the outcomes and processes that will be the focus of the validation process.

*Note: Some of the issues described in this section are similar to those encountered when initially designing an accountability system. However, the purpose of this section is to provide guidance in the evaluation of accountability systems, not for their design. Readers are referred for excellent guidance on the latter to *Designing Accountability Systems: Towards a Framework and a Process* (Gong, 2002), published by CCSSO.*

Major questions

- ♦ What are the goals this accountability system is meant to achieve?
- ♦ Who is to be held accountable for these goals in this system?
- ♦ What indicators are used to represent performance in relation to these goals?
- ♦ How and when are decisions made regarding performance toward the goals?
- ♦ What consequences are associated with different levels of performance?
- ♦ What changes are these consequences meant to effect?
- ♦ How are the intended changes thought to be related to the overall goals?

The following is a quote in which the subject has been intentionally left blank:

“...are intended to have an impact on...the implemented curriculum; the instructional content and strategies; the content and format of classroom assessments; student, teacher, and administrator motivation and effort; the improvement of learning for all students; the nature of professional development support; teacher participation in the administration, development, and scoring of the assessment; student teacher, administrator, and public awareness and beliefs about the assessment, criteria for judging performance, and the use of the assessment results; and the use and nature of test preparation materials...”

(Lane, Park, & Stone, 1998, p. 25)

The actual subject of this quote is “statewide assessment programs,” but one could just as easily fill in that blank with “accountability systems.” In fact, in this era of standards-based reform, accountability systems are usually meant to be the impetus by which assessment results, in combination with other summary indicators, ultimately result in improved student achievement by effecting change in areas of educational functioning like those listed above.

Standards-based accountability systems are supposed to work by applying specific criteria to indicators (e.g., test scores, graduation rates, attendance rates), categorizing each school and LEA as, for example, “excellent” or “needs improvement” based on whether the school or LEA met the criteria, and then assigning consequences corresponding to each category. A state’s academic standards should provide the basis for accountability expectations. The consequences—and even just the category labels themselves—are meant to effect change in areas including those listed above in order to improve specific educational outcomes. Once the consequences have been implemented, the outcomes are re-evaluated to determine whether the intended goal(s) has been achieved.

That is essentially the grand Theory of Action (ToA) for all accountability systems. To tailor this ToA to an individual accountability system first requires answering the following types of questions (recalling the simple functional model of accountability systems from Part I).

TABLE 3. GRAND THEORY OF ACTION QUESTIONS BY COMPONENT

Component	Grand Theory of Action Questions
Goals	What are the goals that this accountability system is intended to achieve? Who is to be held accountable for these goals in this system?
Consequences	What consequences are associated with different levels of performance? What changes are these consequences meant to effect? How are the intended changes thought to be related to the overall goals?
Decision rules	How and when are decisions made regarding performance toward the goals?
Performance indicators	What indicators are used to represent performance in relation to the goals?

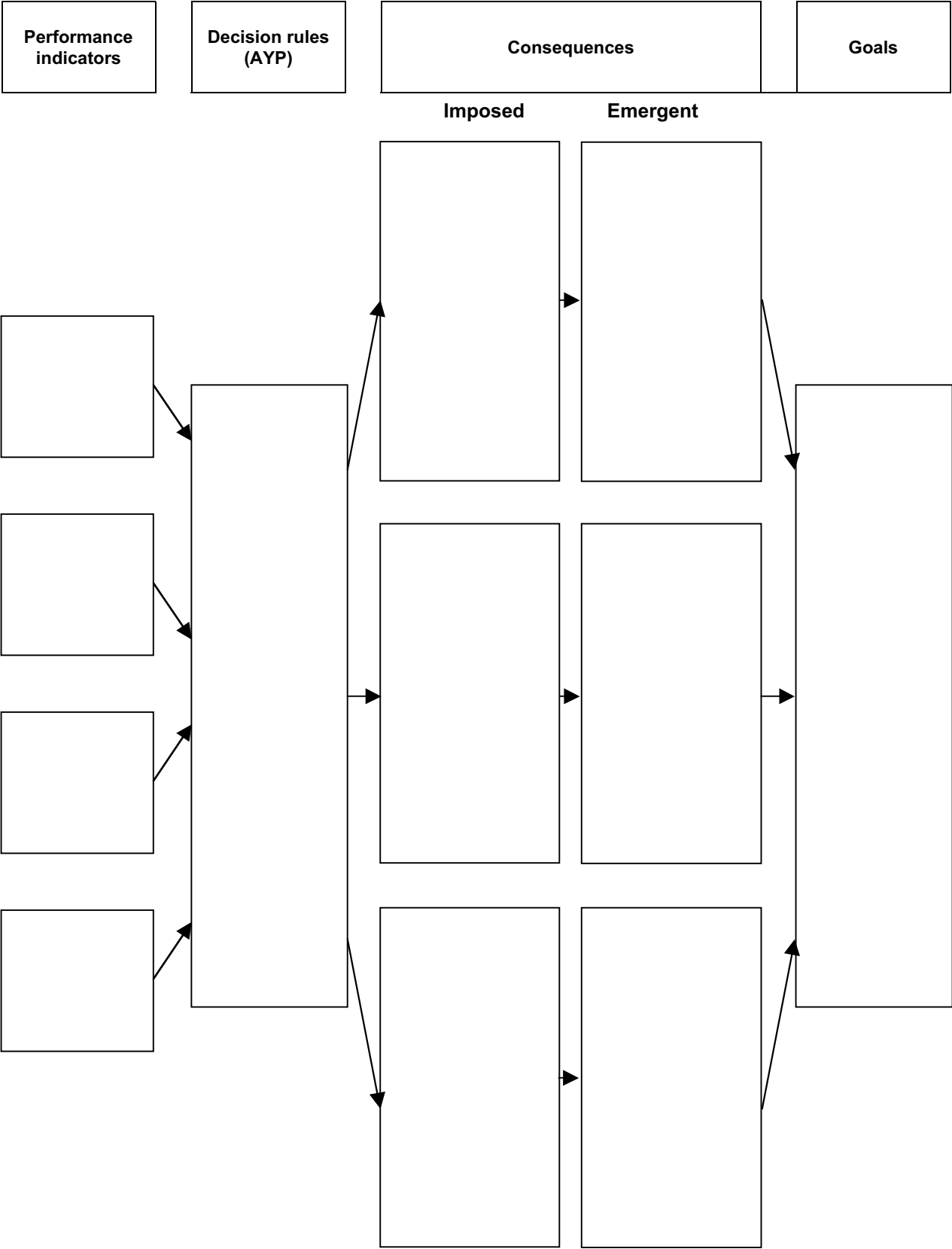
Answers to these questions should address all accountability goals, indicators, decisions, and consequences, not just those required by NCLB.

An outline, such as the one that appears on the following page can be used to organize answers to these questions. Then, a graphic model of the system can be developed to help guide evaluation plans; a blank example of such a model, which could be filled in according to how the questions in the outline were answered, appears after Table 4. Both the outline and the model may differ by content area and/or grade level.

TABLE 4. WORKSHEET FOR INITIAL MAPPING OF MAJOR COMPONENTS AND THE THEORY OF ACTION

Element	Performance indicators	Decision rules (AYP)	Consequences	Goals
Questions	What indicators are used to represent performance in relation to the goals?	How and when are decisions made regarding performance toward the goals?	What consequences are associated with different levels of performance? What changes are these consequences meant to effect? How are the intended changes thought to be related to the overall goals?	What goals is this accountability system intended to achieve? Who is to be held accountable for these goals in this system?
Worksheet	Goal 1 indicators:	To qualify for a reward: When and how often?	Rewards include: Which are meant to:	Goal 1: Who is accountable?
	Goal 2 indicators:	To qualify for sanction: When and how often?	Sanctions include: Which are meant to:	Goal 2: Who is accountable?
	Goal 3 indicators:	To qualify for intervention: When and how often?	Interventions include: Which are meant to:	Goal 3: Who is accountable?

FIGURE 5. SAMPLE MODEL FOR MAPPING THE RELATIONSHIPS AMONG COMPONENTS



The graphic model of the ToA can be an excellent tool for designing and representing the validation process. It is possible to distinguish a number of boxes and a number of arrows (numbers will vary across states and possible content areas and grade levels within states): every box and every arrow must be evaluated as part of the validation process. The questions that correspond to each of these evaluations will be addressed in subsequent sections.

Once the grand ToA has been specified, it would be important to focus in on the right side of the scheme above and closely examine the details of exactly how the consequences that are imposed on schools and LEA are supposed to work. What assumptions are being made about the process of school improvement, about schools' control over factors that seriously affect AYP results, and about how the continuous raising of the AYP bar affects motivation among faculty in highly challenged schools? These types of issues involve a more finely grained view of the ToA and will be addressed in the fourth section of this part of the paper.

SECTION 2

EVALUATING THE INDICATORS

Purpose

To evaluate the accuracy and meaningfulness of the indicators used in the accountability system in general and the AYP model in particular.

Major questions

- ♦ What indicators are part of the accountability system and how is each used?
- ♦ How well do the definitions of these indicators capture what is intended?
- ♦ How reliable are the indicators used to make high stakes accountability decisions?

INDICATORS IN ACCOUNTABILITY SYSTEMS

Indicators are values that represent constructs. Test scores, for example, are indicators of the construct(s) the test is meant to measure, such as reading comprehension or mathematics problem solving. Graduation rates may be considered indicators of how well a school or school system is preparing its students to meet graduation requirements.

The objectives in selecting indicators for use in an accountability system are to achieve as close a match as possible between the indicators and the system goals (Carlson, 2002; Gong, 2002; Marion et al., 2002) and to ensure that indicators can withstand the “pressure” of being used to make high stakes decisions. Validation of the indicator component of an accountability system involves evaluation of the degree to which these objectives have been met.

States use a variety of indicators in their accountability systems. NCLB prescribes the definition and use of some indicators in states' accountability systems. These are summarized in the following table, in addition to a comparison with similar reporting requirements for school, LEA, and state report cards¹³. It should be noted that the indicators that must be included in AYP do not exactly match those that must be included in the report cards (e.g., percent participating for AYP and percents not participating for report cards—although it is unlikely that a state would be reprimanded by the US Department of Education for using percent participating for both purposes).

Some accountability indicators are “qualified” in that they are based on a restricted data range. For example, the percent proficient indicators are to be based only on scores for students who have been enrolled for a full academic year (FAY) at the time of testing. In this case, the definition of FAY is not an indicator in itself but a qualifier of other indicators. Since it affects the calculation and interpretation of the indicator, it must be considered as part of the validation process.

TABLE 5. NCLB REQUIRED INDICATORS FOR AYP

Topic and Indicator	Use for AYP	Include on Report Cards
Reading/Language Arts and Mathematics Achievement		
Aggregate and disaggregated percents scoring at or above the proficient level	X	X
Aggregate and disaggregated percents scoring <u>in each</u> of the academic achievement levels		X
Participation Rates		
Aggregate and disaggregated percents participating in reading/language arts and mathematics assessments	X	
Aggregate and disaggregated percents of <u>non-participating</u> students for the reading/language arts and mathematics assessments		X
Other Indicators		
“Aggregate information on any other indicators...used to determine AYP”	X	X
Disaggregated information on any other AYP indicators	only for safe-harbor	
Aggregate graduation rate for secondary schools	X	X
Disaggregated graduation rate for secondary schools	only for safe-harbor	X

States are not entirely constrained by the list of NCLB-defined indicators. They can include other academic indicators in the accountability systems either by using them as the “other academic indicator” for AYP or by using them to moderate or otherwise inform accountability decisions (Palmer & Coleman, 2003). Some indicators may be used as part of accountability decisions, some to gauge intermediate or process variables, and some simply publicly reported.

Further, state-selected indicators can take many forms. For their other academic indicator at the elementary/middle school level, most states chose to use such

¹³ Since the focus here is on indicators used for accountability decisions, this table provides only an incomplete list of indicators required on school, LEA, and state report cards.

indicators as attendance rates, the percent proficient in science or writing, or changes in the reading and mathematics score distributions not reflected by the percent proficient. At least one state offers LEAs a menu of indicator options from which LEAs could choose. It should be noted that, under NCLB, performance on any of these additional indicators cannot compensate for performance on the prescribed indicators. In other words, other indicators can be used to increase—but not to decrease—the number of schools identified for improvement.

When additional indicators are used beyond AYP, the range of options expands even further. As one example, states can report a variety of indicators and summarize across these indicators as part of the accountability decision (e.g., a school must meet criteria for at least 17 of 22 indicators, in addition to making AYP, to qualify for “excellent” status overall).

TABLE 6. OPTIONS FOR USING STATE-SELECTED INDICATORS IN ACCOUNTABILITY SYSTEMS

Options for state-selected indicators	Accountability decision based on:	How <u>non-AYP</u> state-selected indicators are used:
Other AYP indicator only	AYP indicators	n/a
Other AYP indicator + additional indicators	AYP indicators	State-selected indicators are reported publicly but not used in making formal accountability decisions. Still, they are part of informal accountability in that they— - contextualize the AYP outcomes by inclusion on report cards - provide additional information to better reflect state’s values
Other AYP indicator + additional indicators	Combination of AYP indicators and other indicators	State-selected indicators are used to moderate the accountability decisions of which AYP is a part. Usually, this is a conjunctive relationship, meaning that a school would have to meet AYP criteria <u>and</u> meet additional criteria.
Other AYP indicator + additional indicators	Combination of AYP indicators and other indicators	State-selected indicators are used to make accountability decisions that are separate from AYP (not conditional on AYP outcomes). Separate systems for accountability and accreditation fall into this category.

Generally, the indicators used in accountability systems fall into three categories, based on their data source:

- ♦ Test-based indicators
- ♦ Rates
- ♦ Survey indicators

For validation purposes, the primary focus should be on the indicators used to make accountability decisions, regardless of whether they are prescribed by NCLB or part of AYP. For the present, evaluation of other indicators—even those used in accountability reporting—may need to wait. Survey indicators are generally not used for AYP purposes, so in this paper, only test-based and rate indicators will be considered.

It may be helpful—and will certainly facilitate validation work—to list and define each of the indicators used in accountability decisions as a first step in this process. Such a list could help to clarify what each indicator is meant to

represent, where the data come from, and how the data are translated into the indicator. Table 7 is an example of how this list might look.

EVALUATING TEST-BASED INDICATORS

As described in Part I of this paper, much attention has been focused on validity and the validation process for tests and assessments; there is a relatively solid foundation to guide states in conducting this work. Accordingly, it is not necessary or possible to provide a comprehensive guide to assessment validation in this paper. However, some major questions that are relevant to this work are provided in the discussion that follows. These questions should be addressed as part of a validation process to ensure that the assessment-based indicators that are used in an accountability system are meaningful and appropriate for that use.

All states must use percent proficient indicators for reading and mathematics as part of their AYP decisions. Although some states also use other indicators, the focus here will be on the percent proficient indicators. It should be noted that percent proficient indicators are based on tests, but they are also rates (proportions of populations that meet certain criteria). Therefore, the information in the section following this one, which focuses on rate indicators, should be considered in addition to what appears below in this section.

The primary validity issues with regard to these assessment-based indicators are whether the indicators represent what they are meant to and are appropriate for use in high stakes decision-making. The evidence for consideration of these issues can be organized around two central questions, both of which should be addressed in a validation process:

- ♦ Does the indicator reflect the content and construct(s) it is meant to reflect?
- ♦ How reliable is the indicator?

Two critical points are worth mentioning here. First, recall from the discussion in Part I that validity relates not to the test but to the interpretation and use of test scores. States and testing contractors may be able to offer up proof that they have conducted validation studies with regard to their tests for years.

TABLE 7. SAMPLE LISTING OF INDICATORS USED IN ACCOUNTABILITY DECISIONS

(Note: definitions and data sources are examples only)

Indicator	Definitions	Process and Data source
Percent Proficient: Reading	Conceptual: Each year, the proportion of students in the tested grades whose reading knowledge, skills, and abilities are at or above the level considered proficient for their grade. Over time, as used in NCLB AYP, increases in the percent proficient are assumed to reflect increases in the effectiveness of a school in helping students at a given grade level to attain proficiency in reading. Operational: For each school and LEA, the number of students scoring at or above the proficient level <u>divided</u> by the number of students in the tested grades who have been enrolled from Sept. 30 of the current school year through the last day of the testing window.	Indicator calculated by SDE staff and submitted electronically to LEAs for review by July 1. LEAs request corrections by July 15, with final state response by July 29. Grades 3, 5, 7 – Numerator: Number scoring in the Proficient or Advanced levels on the State Reading CRT (March 15 admin; May 15 data file) Denominator: Total enrollment from matched enrollment file for Sept. 30 to March 15 High School – Numerator: EOC English I. (Total pass from first attempt in 10th grade through March 22 admin of junior year; May 8 data file) <u>minus</u> the number of students who passed the test but were no longer enrolled on the test date Denominator: Total enrollment from matched enrollment file for Sept. 30 to March 22
Participation Rate: Reading	Conceptual: The proportion of all students in the tested grades who participated in the reading assessments used for accountability. Operational: For each school and LEA, the number of students who attempted the reading assessment by responding to at least 10 items <u>divided</u> by the number of students in the tested grades who were enrolled at the mid-point of the testing window.	Indicator calculated by SDE staff and submitted electronically to LEAs for review by July 1. LEAs request corrections by July 15, with final state response by July 29. Grades 3, 5, 7 – Numerator: Number of students who attempted State Reading CRT (March 15 admin; May 15 data file) Denominator: Total enrollment from matched enrollment file for Sept. 30 to March 15 High School – Numerator: Number of students who attempted the EOC English I through the March 22 admin (May 8 data file) <u>minus</u> the number of students who passed the test but were no longer enrolled on the test date Denominator: Total enrollment from matched enrollment file for Sept. 30 to March 22
Attendance	Conceptual: The proportion of the student body that is present on any given school day. Operational: For each school and LEA, the number of students present for each school day, summed across all school days <u>divided</u> by the number of days in the school year.	Indicator calculated by SDE staff and submitted electronically to LEAs for review by July 1. LEAs request corrections by July 15, with final state response by July 29. Numerator: The sum of the number of students present <u>divided</u> by the number of students enrolled across the full school year. Denominator: The number of school days in the school year.

...use of a test score for high stakes accountability purposes may be new and this requires another look at assessment validity issues.

However, use of a test score for high stakes accountability purposes may be new and this requires another look at assessment validity issues. For example, when high stakes are attached to test scores, the motivation to cheat may increase (Harrington-Lueker, 2000; Zlatos, 1996). If a state does not take a close look at its test security practices

and strategies for identifying potential breaches, it may forfeit the meaning of its test scores and their usefulness for accountability purposes—perhaps without even realizing it.

Second, it is the responsibility of the agency that imposes a test and bases decisions, even in part, on test scores to ensure that the evidence supports such uses of these scores. This responsibility does not lie with the testing contractor. However, the agency (e.g., the state department of education) can and should build into its contract with the testing company specifications for conducting validation studies and for providing data to the agency for further validity investigations.

CONTENT AND CONSTRUCT QUESTIONS

Whether and how well an assessment-based indicator reflects what it is meant to reflect depends on how well the assessment—

- ♦ aligns with the intended content domain and not other content domains;
- ♦ engages the intended cognitive and behavioral response processes and not other processes; and
- ♦ minimizes the biasing effects of irrelevant contextual or personal characteristics on test performance.

Alignment

Alignment has to do with the relationship between a test and the content domain that it is intended to measure. States must develop a coherent approach for ensuring alignment between each of its assessments, or combination of assessments, and the Academic Content Standards and Academic Achievement Standards the assessment system is designed to measure. This approach should be initiated during planning and development phases of a test's life cycle and must be monitored with the construction of each test form and revisited each time the academic content and achievement standards are revised.

States that develop custom assessments usually take great care in the alignment of these assessments to their standards during the development phase. This involves the legitimate participation of local, independent individuals (e.g., educators, business people, parents) who understand the standards in establishing the test design and in reviews of test items.

In addition to this development work, however, states that develop custom assessments should conduct the same type of separate, independent evaluation of the alignment between their standards and assessments that states who adopt existing assessments should conduct. The purpose of this independent evaluation is to provide a second opinion on the quality of the alignment from individuals and/or organizations that do not have a vested interest in the standards, the assessments, or the outcomes of the accountability system. Usually, reports from these external evaluations provide states with feedback to use in improving alignment.

When tests are being used operationally, alignment is still a concern. Each test form should comply with the established test design, which helps to ensure that the scores it yields carry the same meaning across time. States should have a system in place for monitoring test form construction. This system should include, but not be limited to, requirements to which contractors must adhere during the construction process.

The following questions should be addressed as part of the state’s approach to ensuring alignment between its standards and assessments:

- ♦ Are the assessments and the standards aligned **comprehensively**, meaning that the assessments reflect the full **range** and **depth** of the standards? Are the assessments as cognitively challenging as the standards?
- ♦ Are the assessments and the standards aligned in terms of **content** and **process**, meaning that the assessments measure what the standards state students should both know and be able to do?
- ♦ Do the assessments reflect the same **degree and pattern of emphasis** on the content as are reflected in the state’s Academic Content Standards?
- ♦ Do the assessments yield scores that reflect the full range of achievement implied by Academic Achievement Standards?
- ♦ Do the assessments measure the knowledge and skills described in its Academic Content Standards and not knowledge, skills, or other characteristics that are not specified in the Academic Content Standards?
- ♦ How are **gaps and weaknesses** identified and what is done to improve the alignment of its standards and assessments?
- ♦ Does the level of performance that has been designated “proficient” for the purposes of accountability reflect developmentally-appropriate expectations for the knowledge, skills, and abilities in each content area as well as the intended uses of the percent proficient indicator?
- ♦ What measures have been taken to minimize or eliminate questionable test preparation practices or breaches of test security that threaten score meaning?

If the state’s assessment system includes assessments developed or adopted at the local level and uses scores from these local assessments in statewide accountability decisions, the state must also gather evidence that these local assessments—

- ♦ are aligned with the state’s Academic Content and Achievement Standards, using the types of alignment questions listed above;
- ♦ are equivalent to one another in terms of content, difficulty, and quality;
- ♦ yield comparable results for all subgroups; and
- ♦ yield results that can be aggregated with those from other local assessments and with any statewide assessments.

Response Processes

Responding to achievement test questions involves engagement of cognitive processes, such as recognition, recall, extrapolation, or deductive reasoning. Responding also involves behaviors such as circling answer choices; filling in blanks, bubbles, or grids; drawing pictures; handling lab equipment; and writing short or extended responses by hand or on the computer. The following questions should be addressed regarding these response processes:

- ♦ Are the intended cognitive processes, and not other processes, being tapped by assessment items? (See the questions in the bias section, below.)
- ♦ Are the behaviors required to record responses developmentally appropriate?
- ♦ Are the answer documents designed to facilitate responses and minimize error during the recording process?

Comparability

States must ensure that test scores carry the same meaning and support the same interpretations across students. This is especially important when scores for subgroups of students, such as students with disabilities or students who are not yet proficient in English, are held to the same criteria as scores for the total student population and used to make decisions with high stakes consequences.

The effects of irrelevant contextual and personal characteristics can inhibit the interpretability of test scores. For example, if the lone prompt on a writing assessment mentioned exploration of an attic, this could confuse some young students who have always lived in apartments and have no idea what an attic is. More broadly, if a science assessment requires manipulation of lab equipment that a particular school does not have, the students in that school would be at a disadvantage simply because they attend that school. Accordingly, several issues must be addressed:

- ♦ Are the test and item scores correlated with outside variables as intended (e.g., scores are correlated strongly with relevant measures of academic achievement and weakly correlated, if at all, with irrelevant characteristics, such as demographics)?
- ♦ Do all students in the tested population have essentially equivalent opportunities to learn the material in the standards?
- ♦ With regard to the fairness and accessibility of the assessment system for all students, including students with disabilities and students with limited English proficiency—
 - Are administration conditions essentially equivalent across students and schools, meaning that the conditions afford all students the opportunity to demonstrate what they know and can do without undue advantage or disadvantage?
 - Do scores that are based on accommodated administration conditions allow for valid inferences about these students' knowledge and skills and can they be combined meaningfully with scores from non-accommodated administration conditions?

- Do the alternate assessments yield scores that are interpretable with regard to the target constructs?
- How were fairness and accessibility considered during the development of the assessment items?
- What evidence supports the comparability of assessments where comparisons are made between groups or over time?
- Are appropriate accommodations available to students with disabilities and limited English proficient students and are these accommodations used reasonably and as necessary to yield accurate and reliable information about what students with disabilities and limited English proficient students know and can do?
- Are the guidelines for inclusion clearly communicated across the state? Are these guidelines understood and followed well enough so that inclusion practices can be considered standardized across the state?
- If the state allows LEAs enrolling limited English proficient students to select their own assessments of English proficiency, how is the state ensuring the quality and comparability of these assessments?

RELIABILITY

With regard to the issue of **reliability** as described in the *Standards* (AERA/APA/NCME, 1999), each of the following questions should be addressed for each type of test score used for accountability purposes:

- ♦ Is the estimated reliability of each type of reported test score (e.g., percent proficient) adequate for the intended uses of the test scores, based on data for the reported student population and each reported subpopulation?
- ♦ Is score reliability effectively communicated to those who use the scores?
- ♦ Is the conditional standard error of measurement and decision-consistency at each cut score specified in its Academic Achievement Standards—and especially at the proficient vs. non-proficient demarcation—adequate to support classifications?
- ♦ What evidence supports the generalizability of test score interpretations for all relevant sources, such as persons, items, schools, forms, and raters?

States (or LEAs) should require their assessment contractors to provide some of the information needed to answer these questions. For example, testing contractors should provide reliability estimates for each reported assessment score, including those for disaggregated student groups, as well as the conditional standard error of measurement at each relevant cut score. Contractors may also be willing to support analyses related to decision consistency and generalizability. States should recognize that consideration of these issues has implications for data collection and, thus, must be discussed with a contractor several months prior to test administration or field testing.

EVALUATING RATE INDICATORS

Rate indicators are proportions of a given population that meet a specific criterion or set of criteria; percent proficient, participation, and graduation rates are all rate indicators required for NCLB AYP. A number of states also use attendance rates as the other indicator for elementary and middle schools. In this section, the discussion will focus primarily on rates other than the percent proficient. Although this indicator is a rate, it is also a summary test score. Therefore, it should be evaluated as part of the assessment validation process. Of primary concern would be consideration of the conditional standard error at the cut point between the proficient level and the level below it and the decision-consistency with which students are classified as proficient vs. not proficient.

In their simplest form, rates are calculated by assigning every member of a population a “1” if they met a criterion and a “0” if they did not, summing these ones and zeros, and then dividing by the total number of the target population.

To evaluate the quality of rates, it is necessary to examine the extent to which the definitions for the numerator and the denominator match the practices used to generate those figures. In addition, the reliability of the rate should be estimated. Thus, there are two primary questions for rate indicators:

- ♦ How well do the practices for generating rates align with the definitions for the numerator and denominator?
- ♦ How reliable are the rate indicators?

ALIGNMENT OF PRACTICE WITH DEFINITIONS

For rate indicators to support valid interpretations, they must reflect what they are meant to reflect as accurately as possible. Evaluation of accuracy requires a comparison between the definition of the indicator and the way in which it is actually calculated.

Calculation of rates involves three basic steps:

1. Compilation of student-level records;
2. Calculation of summary counts at the school-, LEA-, and state-levels as appropriate; and
3. Calculation of the rate by dividing one count by another, as specified in the rate definition.

TABLE 8. VALIDATION QUESTIONS AND STRATEGIES RELATED TO DATA

When state enters process	Primary validation questions	Examination and prevention strategies
1. State receives student-level data	Issue: must evaluate how LEAs/schools gather and code data and the processes for retrieving and analyzing those data at the state Student-level records How and by whom are student records maintained and updated? When and how can LEAs make corrections to student-level data as part of the accountability decision process? Summaries How are counts generated and by what means are they verified? Rates How are rates generated and by what means are they verified?	• Annually sample and audit LEA (and possible school) data collection and management methods • Implement a single data/summary/rate review and correction process with a strict completion deadline that will not allow delays to accountability decisions • Review and monitor internal data analysis processes
2. State receives summary counts	Issue: must evaluate how LEAs/schools gather, code, and summarize data and the processes for retrieving and analyzing those data at the state Student-level records What evidence supports the accuracy and legitimacy of the student-level records? Summaries What evidence supports the accuracy and legitimacy of the summary counts? Rates How are rates generated and by what means are they verified?	• Annually sample and audit LEA (and possible school) data collection and management methods • Annually sample and audit LEA methods for summarizing data • Implement a review process for the final rate indicators with a strict completion deadline that will not allow delays to accountability decisions • Review and monitor internal data analysis processes
3. State receives indices from schools or LEAs	Issue: must evaluate how LEAs/schools gather, code, summarize, and analyze data and the processes for retrieving those data at the state Student-level records What evidence supports the accuracy and legitimacy of the student-level records? Summaries What evidence supports the accuracy and legitimacy of the summary counts? Rates What evidence supports the accuracy and legitimacy of the rates?	• Annually sample and audit LEA (and possible school) data collection and management methods • Annually sample and audit LEA methods for summarizing and analyzing data • Implement a process for submitting rate indicators with a strict completion deadline that will not allow delays to accountability decisions • Review and monitor internal data analysis processes

Specific practices for generating rate indicators vary across states and across indicators within states, but generally take one of the following forms:

- ♦ LEAs/schools or the testing contractor transmits student-level data to the state via a statewide pupil database and the state calculates the indices;
- ♦ LEAs/schools report summary data (e.g., counts) to the state and the state calculates the indices; or
- ♦ LEAs/schools calculate their own indices and report these to the state.

Depending on where the state enters the process, validation will require different types of examinations.

The table below provides examples of some problematic data practices one state uncovered in its studies and the strategies the state is using to resolve them.

TABLE 9. EXAMPLES OF VALIDITY-THREATENING DATA PROBLEMS AND THEIR RESOLUTIONS

Problem	Resolution
Students being excluded from accountability system by manipulating authorized reporting codes in the Student Information System	Data audits of highly suspect and over-reporting of key exclusion codes (e.g., moved out of state)
Manipulation of standardized administration of statewide assessments	Public can report irregularities to the Department using a web-based system. Data validation unit conducts case-by-case investigation.
Manipulation of dropout and attendance data	Data audit of entire population to investigate extreme status and/or status conducted by the Data Validation Unit. Conduct on-site audit of attendance and dropout records.
Reconfiguration of schools to meet appeals requirement for temporary exclusion	Policy requirements place rewards or corrective action status on new school if 50% of students go to another school.
Students with disabilities are receiving educational services at a centralized area (school). This results in disproportional SWD subgroup size and in effect reducing the burden of the “home school” and increasing the burden of the “centralized service” school.	Student performance and behavioral data are routed back to the “home” school and aggregated into that school’s totals.
Special education services, through the encouragement of the IEP committee, agree to terminate services for targeted disabilities (e.g., speech and language, learning handicapped) and use other classifications to serve needs (e.g., 504, migrant, at-risk, alternative programs), thus reducing subgroup membership.	SDE conducts “change of status” audits during local educational agency IDEA compliance reviews. Require, in local policies, language explaining rationale for status change with central office oversight.
Students are advanced across grades (grade skipped) with “high stakes” requirements using the IEP process.	SDE conducts student-to-student matches from the prior year to identify individuals whose grade placement change is >1 level. Require local pupil progression plans outline policies restricting practice and mechanisms for monitoring during the school year.

In addition to these kinds of examinations of data quality, states should take care in developing the directions provided to schools and LEAs as part of data coding—including those directions that are part of the assessment package. For example, every state that is implementing accountability decisions has had to carefully consider who qualifies as a student in a given school. Certainly, the student who regularly attends core classes qualifies. However, the state must establish clear guidelines for how to account for other students, such as those who attend classes in out-placement facilities, are migrant, or are unable to be physically present at the school due to chronic medical conditions. The state will also have to evaluate the degree to which these guidelines are followed in practice.

Equally challenging will be the evaluation of the application of the “numerator criteria.” For example, exactly what does a student have to do to qualify as having participated in an assessment? Does the student have to do more than complete the demographic page? Does the student have to attempt at least a certain minimum number of items on the test? If the assessment is composed of multiple tests, administered at different points in the school year (e.g., a reading test and a separate writing test which, together, comprise the reading/language

arts assessment), does the student have to attempt only one test or both to count as having participated? These are among the types of policy decisions a state must make as part of implementing a high stakes accountability system.

RELIABILITY

For rate indicators, reliability can be thought of in two ways. First, one could ask how well the rate represents the “true” proportion that would be observed “given a sample of all possible students who could attend that school” or LEA. Alternatively, one could ask whether the observed proportion is “different than the target proportion” (Marion et al., 2002, p. 66).

The first question can be addressed by estimating the standard error of the proportion (SE_p) using the following calculation:

$$SE_p = \sqrt{\frac{\text{observed proportion} (1 - \text{observed proportion})}{N}}$$

For example: If a school has 100 students enrolled in the tested grades and 93 of these students participated in the reading assessment, the participation rate would be $93/100 = 93\%$.

To calculate the standard error for this proportion, perform the following:

1. Transform 93% into decimal form (= .93)
2. Subtract .93 from 1 (= .07)
3. Multiple .93 and .07 (= .065)
4. Divide .065 by the total number of students (= $.065/100 = .00065$)
5. Find the square root of .00065 (= .025)

The SE_p (.025 in the above example) is then used to create a confidence interval around the observed proportion. This confidence interval has a lower limit, which is calculated by subtracting the SE_p from the observed proportion, and an upper limit, which is calculated by adding the SE_p to the observed proportion.

For the example, this would mean the lower limit = $.93 - .025 = .905$ and the upper limit = $.93 + .025 = .955$. Thus, the “true” proportion for the school would be somewhere between .905 and .955.

The second question, whether the observed proportion is different from the target proportion, may be more appropriate for rates with a specific AYP target. For example, the participation rate target is 95%. Using the information from the example above, one could ask whether the observed proportion of .93 really differs enough from .95 to warrant an AYP failure in this area.

Certainly, one could use the SE_p as evidence that the true proportion for the school could have been .95 or slightly higher, but a more refined test may be more appropriate given the high stakes associated with failure to reach the target in this case.

The following process can be used to determine whether the observed proportion is lower than the target proportion (see Marion et al., 2002, for an extended discussion of these issues).

1. Determine how certain you wish to be about whether schools' observed proportions are lower than the target. In doing so, realize that—
 - a. common choices are 95%, 98%, or 99%; 100% is not realistic—it would essentially mean that all schools would always make AYP.
 - b. the higher the level of certainty, the more sure you are that schools with true scores above the target will make AYP. At the

...the cost of increased reliability may be decreased validity.

same time, as the level of certainty increases, so does probability that schools with true scores below the target will make AYP. In other words, the cost of increased reliability may be decreased validity.

2. Calculate the statistic necessary for this test. In this case, it is a z-statistic as calculated below:

$$Z = \frac{\text{observed proportion} - \text{target proportion}}{\sqrt{\frac{\text{target proportion} (1 - \text{target proportion})}{N}}}$$

3. Then, find a basic statistics book and locate the table in the appendix that provides the critical values of the z-statistic for different levels of certainty. For 95% certainty, the z-statistic critical value is 1.645. For 98% it is 2.054 and for 99% it is 2.326.

Compare the number that results from the above equation to the critical value for the level of certainty you have selected. If the absolute value from the above equation is less than the critical value, then the observed proportion is not significantly different from the target value.

Using the numbers from the example on the previous page, the absolute value of the z-statistic would be 0.9. Since this is less than the critical values noted above, one would be 99% confident that the observed participation rate of .93 is not significantly different from the .95 target.

SECTION 3

EVALUATING THE DECISION RULES

Purpose

To evaluate the reliability and accuracy of accountability decisions, including AYP.

Major questions

- ♦ Do the results of the AYP model support the goals of the accountability system?
 - ♦ Were the “right schools” identified for rewards, sanctions, and interventions?
 - ♦ Are the results of the AYP model stable over time?
-

INTRODUCTION

It is important to distinguish between the validity of the AYP model and the validity of the state’s overall accountability system. The overall statewide accountability system includes other components including the assessment system, the system of rewards and sanctions, the reporting system, and the technical assistance and professional development programs. The purpose of the overall statewide accountability system may be to increase the academic achievement of all students in the state. A critical component of this system is the decision-making process that, under NCLB, includes AYP. When one asks, “Were the ‘right’ schools identified for improvement,” as Edie and her colleagues did in the scenario that opened this paper, they are asking about the reliability and validity of the decision-making process. Suggestions for gathering evidence in relation to this question are the focus of this section.

This section has two parts. First, strategies are offered for evaluating the process by which schools and LEAs are initially identified for improvement. Then, the appeals process that allows schools and LEAs to formally request a reconsideration of their improvement status is addressed.

EVALUATING THE IDENTIFICATION PROCESS

If indicators were selected that truly support the desired purposes of the model and the data collected for those indicators proved to be reliable and accurate, the next issue to address is the degree to which the data interpretation and resulting decisions adequately met the purposes established for the system. For the AYP model specified in NCLB, one of the major purposes is the identification of schools that are meeting the AYP requirements—but there is an implicit generalization that failure to meet AYP (for two or more consecutive years) means that a school or LEA is not effective and needs improvement. Basically, the validity of the model is based on the degree to which the AYP determinations identify the correct schools—those in greatest need of improvement—while not identifying schools that are doing an effective job educating all students.

The AYP model prescribed in NCLB is a classification system. Each school and each LEA is classified annually into one of two groups—met AYP or did not meet AYP. If the model were perfectly reliable and accurate, every school and LEA would be correctly classified. Those needing improvement would be assigned to the “not met” group while those not needing improvement would be assigned to the “met” group.

The AYP model is actually more complicated than indicated above. Since schools/LEAs are considered “in need of improvement” when they have failed to meet AYP for two consecutive years and are considered to “no longer be in need of improvement” when they have met AYP for two consecutive years, the degree to which the model is identifying the correct school (and meeting the major purpose of the AYP model) would require that the identification for improvement and exit from that status would be the most appropriate indicator of the validity of the overall AYP model.

To simplify the process and allow the validity of the system to be judged earlier, we could assume that reliable and accurate annual AYP decisions would be sufficient to indicate that the overall AYP model is valid. For purposes of the following discussion, the following relationships will be used:

- ♦ Meeting AYP = “school/LEA is not in need of improvement”
- ♦ Not meeting AYP = “school/LEA is in need of improvement”—even though the school would not actually be identified for improvement unless AYP is not met for two consecutive years

No evaluation system produces perfectly accurate outcomes. Given a model that assigns schools and LEAs to two groups based on certain criteria, there are four possible outcomes for each decision:

- ♦ School/LEA in need of improvement identified as such (true positive)
- ♦ School/LEA not in need of improvement identified as such (true negative)
- ♦ School not in need of improvement indicated as needing improvement (false positive)
- ♦ School in need of improvement identified as not needing improvement (false negative)

These four outcomes can be illustrated as a four-fold truth table.

FIGURE 6. SAMPLE FOUR-FOLD TRUTH TABLE FOR AYP

		AYP Determination (Results of the AYP Model)	
		Did Not Meet AYP	Met AYP
True School Effectiveness Status	Does Not Need Improvement	False Positive (Error)	True Negative
	Needs Improvement	True Positive	False Negative (Error)

If the AYP determinations were perfectly accurate, the true positive and true negative cells (those shaded in the table above) would contain the numbers of schools/LEAs actually needing or not needing improvement (as defined earlier). The false positive and false negative cells would both contain zeros.

A simple measure of the accuracy of the decisions made using the model would be the sum of the numbers in the “true” cells divided by the total number of schools or LEAs for which AYP determinations were made. In the case of a perfectly reliable model, the value would be 1.00 indicating that 100% of the schools or LEAs were correctly classified.

Of course, the identification of a suitable external criterion that allows one to determine the “true” AYP status for each school is a major challenge¹⁴. Given such a criterion measure, the validity of the AYP model could be determined using the simple procedure described above.

Another complicating factor is that the AYP determination made using the AYP model prescribed by NCLB is based on a conjunctive (non-compensatory) set of standards. Failure to meet the standard on only one of many standards results in the AYP determination for the school or LEA being “not met.” The use of conjunctive standards “usually makes the accountability system much less reliable” (Gong, 2002, p. 9) and is seen as a common error in the design of effective accountability systems (Hill, 2000).

Obviously, the simple “met/not met” dichotomy cannot truly represent the degree to which a particular school or LEA needs improvement. It has been argued that such a system might routinely classify schools/LEAs as needing improvement from year to year (Kane & Staiger, 2002; Marion et al., 2002, p. 86).

...a system might routinely classify schools/LEAs as needing improvement simply by chance due to the number of conjunctive standards that must be met from year to year (Kane & Staiger, 2002; Marion et al., 2002, p. 86).

DETERMINING THE LEVEL OF NEED FOR A SCHOOL

Assuming that a state’s AYP model is valid, the “need for improvement” would be very different in a school where only one subgroup missed the annual measurable objective (AMO) in one subject by a few percentage points and a school where many subgroups missed the AMO in both subjects by a large margin.

While the AYP model in NCLB makes no distinction between the two hypothetical schools above (the same sanctions must be applied to both schools), some sort of scale could be developed to represent the degree to which the school failed to meet the criteria in the AYP model. A ratio could be calculated by dividing the number of separate NCLB standards missed (e.g., students with disabilities, elementary grade span, mathematics) by the total number of standards for which the school was held accountable. Of course, this would work best if many of the NCLB subgroups counted for AYP at the school; and may not work well at all if most schools only have two or three subgroups for which they are held accountable. In the latter case, there would be virtually no variance on which to construct a useable numeric scale.

For a school that missed one of 10 possible standards, the ratio would be 1/10 or .10. For a school that missed 4 out of 5 possible standards, the ratio would be 4/5 or .80. The schools or LEAs could be ranked by this ratio—schools with the

¹⁴ Personal communication, Dale Carlson, September 16, 2003.

largest ratio values could be considered to be in greatest need of improvement. This “measure,” of course, would be very rough and would be subject to limitations such as the fact that different schools would have different numbers of possible standards, different numbers of students in the various subgroups, and differences in the magnitude by which subgroups missed the AMOs.

Depending on the degree to which the AMO measures are comparable across subject areas and grade spans, the magnitude by which each subgroup missed the AMO could be used to compute a more sensitive “needs improvement” measure. Given two schools that each missed 3 of 10 standards (simple ratio = $3/10 = .30$), a measure factoring in the degree to which each school missed the AMO could produce different values, reflecting the differences in the need for improvement at the two schools.

An example of factoring in the magnitude by which groups in a school missed the AMO follows.

TABLE 10. EXAMPLE OF DIFFERENCE IN MAGNITUDE OF AMO MISSES

School A	Group	% Proficient	Missed AMO by:
3 groups missed the AMO of 60% proficient	1	30%	30 points
	2	25%	35 points
	3	20%	40 points
Average difference = $(30 + 35 + 40) / 3 = 105 / 3 = 35$			
School B	Group	% Proficient	Missed AMO by:
3 groups missed the AMO of 60% proficient	1	50%	10 points
	2	59%	1 point
	3	50%	10 points
Average difference = $(10 + 1 + 10) / 3 = 21 / 3 = 7$			

The average difference by which School A’s groups missed the AMO was much greater than the average difference for School B. Although each school had the same number of groups missing the AMO, the average differences (35 vs. 7) indicated that School A has more need for improvement.

The example above used an average of the group difference values (% Proficient minus the AMO) as a rough measure of need. That would be appropriate if the number of students in each group missing the AMO was similar in size or if one desired to weight each group equally, regardless of size. The latter is, basically, what the NCLB AYP model does since a school fails to meet AYP if any one subgroup (meeting the state’s minimum n requirement) fails to meet the AYP criteria. To take the size of the group into consideration, the simple averaging procedure can be modified to weight the group differences based on the number of students in the group. In that case, each student counts equally toward the “needs improvement” measure.

An example using a weighted average difference follows.

TABLE 11. COMPARISON OF AVERAGE AND WEIGHTED AVERAGE DIFFERENCES IN GROUP AMO MISSES

School A (3 groups missed the AMO of 60% proficient)				
Group	% Proficient	Missed AMO by:	N	Weighted Differences (N x # Points Missing AMO)
1	30%	30 points	200	200 X 30 = 6,000
2	25%	35 points	200	200 X 35 = 7,000
3	20%	40 points	40	40 X 40 = 1,600
Sum = 105			Sum = 440	Sum = 14,600
Average Difference = $105 / 3 = 35$			Weighted Average Difference = 33.18	

The example above is for the same school used in the earlier example for the un-weighted average difference method. The un-weighted average difference was 35 and the weighted difference is 33. The weighted difference value is lower because, although Group 3 missed the AMO by the widest margin (40 points), there were far fewer students in that group than in Groups 1 and 2.

Using the same n-counts for Groups 1, 2, and 3 in School B, the weighted average difference value for that school is **5.91**. Again, the data indicate that School A, with a weighted difference of 33, has more need for improvement than School B, with a weighted difference of 6.

Using the Measures of Need

The major purpose of this section is to explore some methods states can use to demonstrate that their AYP model results are valid for the intended purposes. The rather simple methods described earlier produce numerical values that can potentially be used for that purpose. There are, however, other uses for these values. Although the sanctions specified in NCLB must be applied to schools failing to meet AYP over consecutive years regardless of the actual level of need (i.e., the measures described earlier cannot be used to change whether a school is or is not identified for improvement, corrective action, or restructuring under NCLB), they can be useful for other purposes. For example, the level of need at schools identified for improvement could be used to determine the kind or level of resources and technical assistance that a state or LEA might provide to a school. That is, a more informed decision could be made regarding the assignment of resources than could be made knowing only whether a school did or did not make AYP (CEP, 2003, p. 41; Marion et al., 2002, p. 15). A LEA audit tool for use by states in determining intensity and type of technical assistance needed by schools is currently being developed by the Comprehensive Assessment Systems (CAS) State Collaborative on Assessment and Student Standards¹⁵.

¹⁵ A project of the Council of Chief State School Officers, Washington, DC

Another use for the level of need measure is determining the effect of various changes or modifications in the state's AYP model. Changes might include small changes such as different alpha levels for confidence intervals or changes in the minimum group n-count or larger changes such as adding a new assessment to the model. This procedure could be valuable both in the initial model development process and as a state makes changes to its model over time. Principle 9 in the Consolidated State Application Workbook (ED, 2002) focuses on the validity and reliability of the statewide accountability system. Critical Element 9.3 requires each state to tell how it has planned for the incorporation of anticipated changes in assessments. Specifically, ED asks if the state has a plan to maintain continuity in AYP decisions necessary for validity through planned assessment changes and other changes necessary to comply fully with NCLB and a plan for periodically reviewing its state accountability system so that unforeseen changes can be quickly addressed. In using measures of need for the purposes above, a state could examine the degree to which those changes affect the measure of need or it could examine the degree to which the changes in the measure of need affect the estimated validity and reliability of the AYP model using procedures discussed later in this section.

The remainder of this section will discuss some ways of using the measures of need to provide evidence regarding the validity (or lack of validity) of the state's AYP model. Another critical element under Principle 9 in the ED Consolidated State Application Workbook deals directly with the validity and reliability of a state's accountability system. Critical Element 9.1 requires a state to tell how its AYP determinations meet the state's standard for acceptable reliability. ED used the following examples as evidence that a state is meeting the requirements of 9.1:

- ♦ State has a defined method for determining an acceptable level of reliability (decision consistency) for AYP decisions.
- ♦ State provides evidence that decision consistency is (1) within the range deemed acceptable to the state, and (2) meets professional standards and practice.
- ♦ State publicly reports the estimate of decision consistency and incorporates it appropriately into accountability decisions.
- ♦ State updates analysis and reporting of decision consistency at appropriate intervals

The examples above indicate that ED appropriately considers the reliability of the accountability system (particularly, the decision consistency of the state's AYP model) to be essential in demonstrating that the accountability system is valid. According to Gong (2002), "states should perform reliability analyses to ascertain that the level of error or uncertainty associated with accountability decisions is acceptable to the [state] and to key policy makers ...states need this type of information for legal and professional defensibility of high-stakes programs" (p. 6).

As discussed at the beginning of this section, validity of the accountability system can be conceptualized at two levels: (1) validity of the entire statewide accountability system (i.e., all components including sanctions, rewards, and improvement strategies) and (2) validity of the AYP model based on the degree to which the AYP determinations identify the correct schools—those in greatest

need of improvement—while not identifying schools that are doing an effective job educating all students.

It is posited that for an AYP model, as for an assessment system, reliability is a necessary (Marion et al., 2002, p. 23), but not (in itself) sufficient, requirement for validity. The remainder of this section will address validity and reliability issues as they relate to accountability models, particularly AYP under NCLB.

VALIDITY OF THE AYP MODEL

The purpose of this paper is not to discuss possible technical issues related to the AYP model mandated under NCLB. It must be recognized, however, that some of those issues are challenging and may affect a state's ability to design and implement a model that exhibits acceptable levels of reliability and validity.

It is interesting that NCLB places strong emphasis on ensuring that states only implement educational programs under Title I and other sections of the Act that are research-based and have been shown to be effective. This is also a requirement in a state's approval of supplemental educational service providers under Title I. This is in contrast to the specific AYP model mandated for use by all states — one on which little, if any, study had been conducted prior to the drafting of NCLB. Work done as the Act was being debated in Congress (Kane, Staiger, & Geppert, 2001; Riddle, 2001) and studies conducted since the Act became law in January 2002 (CEP, 2003; Hill, 2002; Linn, Baker, & Betebenner, 2002) overwhelmingly indicate that there are serious technical issues to be considered. This makes it even more important for states to have some way to judge the reasonableness of the results produced by their AYP models.

As noted earlier, while there is a rich body of literature and a reasonably long history of studies dealing with issues of reliability and validity in the field of assessment, the reliability and validity of accountability systems has only recently begun to be studied. Methods for determining the reliability and validity of such systems have yet to be developed and tested. Most of the recent work on accountability systems (basically, evaluation systems applied to schools and LEAs) has focused on characteristics of models for rating, ranking, and classifying schools. Much of the data used in those studies was generated using Monte Carlo techniques and/or was actual data from various states that have had accountability systems in place (e.g., California, Kentucky, North Carolina, and Texas). Studies and papers by Hill and DePascale (2002; 2003), and Hoffman and Wise (2000) have examined the reliability of school ratings and accountability systems. Those issues are discussed in greater detail under Stability and Reliability in the AYP Model and under Misclassification Error and Validity of an AYP Model. To date, there is no satisfactory way to compute a coefficient of validity, especially of the validity of the overall accountability system (Marion et al., 2002).

Over the last few years, a set of standards has been developed (Baker, Linn, Herman, & Koretz, 2002) describing characteristics of good accountability systems (see Appendix B). Just as the *Standards for Educational and Psychological Testing* (APA/AERA/NCME, 1999) are used to ensure best practice in the development and implementation of assessment systems, a set of standards for accountability systems can provide a basis for conceptualizing best practice for the design and implementation of accountability systems and may help guide the study of reliability and validity in accountability systems.

MEASURES OF NEED AND THE VALIDITY OF AYP MODELS

The measures of need can be helpful to a state in examining the validity of its AYP model. Basically, the information that follows illustrates ways that states may be able to use measures of need along with other information for this purpose. An examination of the various measures of need to determine whether the schools with the highest and lowest values or rankings seem to be the correct schools would illustrate an attempt to establish face validity. Face validity is, however, not a technical type of validity and cannot be used as evidence of the actual validity of the AYP model. All the methods illustrated in this section for examining validity use measures of need along with some other objective information. The relationship between the two variables (or similarities in the pattern of results) is used to help determine the level of construct validity or criterion-related validity. Comparisons of measures of need to the results from another accountability model or school rating system would represent an examination of construct validity. Comparisons of measures of need to direct objective observations of school effectiveness would represent an examination of criterion-related validity.

Before discussing the use of the measures of need for examining the validity of the AYP model, it is important to note the differences in the various measures:

- ♦ The simplest measure, a ratio based on the number of groups in the school that missed AYP looks at the AYP decision for each group exactly as outlined under NCLB. That is, the “met” or “did not meet” determination would incorporate the test participation rate (that must be at least 95%) as well as the “safe-harbor” provision, if applicable. The safe harbor provision allows a group that did not meet the AMO to make AYP if the percentage of non-proficient students were reduced by a criterion amount from the previous year.
- ♦ The weighted and un-weighted differences measures are based only on a comparison of each group’s achievement proficiency to the AMO. Those measures should provide numerical values with more variation than the values produced by the simple “number of groups” measure above. However, they do not include the effect of the test participation rates or the application of the safe harbor provisions to the group level AYP decisions.

There is one other technical issue that affects all of the measures of need discussed in this section. Many state accountability systems or other systems for rating schools have included measures of improvement or growth. Given the large differences in student populations from school to school, or even from year to year within one school, this seems to make sense. Even NCLB contains a provision that allows a group of students that does not meet the AMO to make AYP if the number of non-proficient students decreased by ten percent of the percentage not proficient the previous year. This provision is meant to give credit for improvement in student achievement. The problem with the above “safe harbor” provision is that the improvement in student achievement is based only on the change in the percentage of students reaching the proficient level on the state’s academic assessments. In fact, the entire AYP model is affected by this issue, because, while the Act refers to adequate yearly progress, the determination of whether a group makes AYP is based on the percentage of students in the group that scored proficient, not progress toward that objective.

This issue is raised to illustrate that much of the useful information (and score variance) in a state's assessment program is lost by moving from units such as scale scores to a set of proficiency levels and finally to the dichotomy (proficient or not proficient) required for use under NCLB. Several researchers have examined the problems associated with the use of coarse reporting statistics in accountability systems (Hanushek & Raymond, 2002; Hill, 1997; Hill, 2000; Linn, Baker, & Betebenner, 2002; Thum, 2003).

While proficiency levels are useful for reporting student achievement to parents and the public, using the more comprehensive underlying scales would improve the ability of an accountability model to measure changes in student achievement. Nevertheless, NCLB requires that the AYP model use only percentages of students falling at or above the proficiency cut point—so all the measures of need described in this section are based on those percentages.

Probably the simplest way to use the measures of need for examining the validity of the AYP model is to rank all the schools in the state on the selected measure. Taking only the schools falling at the top and bottom of the rankings (e.g., the top and bottom 10% or the top and bottom 25%), an informed decision could probably be made concerning whether certain schools in the high need and low need groups would be expected to fall in the opposite group. Unless there are only slight differences in schools throughout the state, it is likely that there will be substantial differences in the needs measures of the schools in the high need and low need groups. If some of the schools in the high need group are judged to actually belong in the low need group (and/or some of the schools in the low need group are judged to actually belong in the high need group), the validity of the state's AYP model for identifying schools in need of improvement while not identifying schools that are not in need of improvement would be suspect.

Of course, the process described above calls for some type of informed judgments to be made concerning the schools in the state that are in most (and least) need for improvement. The objectivity and accuracy of such judgments may not be sufficient for the process to provide credible evidence concerning the validity of the AYP model.

Another relatively simple method for examining the validity of the AYP model is available if the state has collected reliable information concerning effective schools and schools where there are problems affecting student achievement. Examples could include data from surveys or site visits. Using only the schools for which the effectiveness data are available, one could determine whether all the effective schools fell in the low need group and all the schools with problems fell in the high need group. If some of the schools were found in the wrong group based on the ranked needs measure, the validity of the AYP model would be questionable.

States that have or had a statewide school level accountability system in place may have a variety of options for using the measures of need for examining similarities and differences between the results of the NCLB AYP model and current or past results on the other accountability model. An assumption would be that both models have the same purpose (or at least very similar purposes). Comparisons of the two models could be conducted in various ways depending on the kind of data the other model produces:

- ♦ If the other model yields some type of numeric score with a relatively wide range (e.g., a 0-100 accountability index) and those data can be

assumed to comprise an interval or ordinal scale, some common statistical measures of relationship could be used. If the assumption of an interval scale is made, the relationship between the results of the two models could be examined by running a Pearson product moment correlation analysis between the value from the state's other model and one of the measures of need described earlier for the state's AYP model. If the correlation coefficient is moderate to high, the models are producing similar results. The degree to which the current or former "other" model was accepted as valid for accountability purposes, a strong relationship might be considered reasonable evidence of the validity of the AYP model.

- ♦ A correlation based on interval scales could also be run between one of the measures of need for the AYP model and the kind of ratings produced by some state accountability models (e.g., a value between 1 and 5), but there probably would be less variance in the school rating values than in an index such as that described above.
- ♦ If the other model yields rankings of schools throughout the state, the relationship between the ranking produced by the other model and one of the measures of need described earlier for the AYP model could be examined by running a rank correlation such as Spearman or Kendall.
- ♦ If the other model yields labels or categories (e.g., excellent, good, fair, poor), a cross-tabulation procedure could be used. First, the measure of need would be used to place schools into different groups by dividing the school distribution into a number of categories. This could be done by using the quartile or quintile points in the school distribution or by simply dividing the score range of the measure of need into equal intervals (e.g., if values on the measure of need range from 10 through 90, "cut points" could be set at 30, 50, and 70 to form four groups of schools). While the categories formed using the measure of need would be based on somewhat arbitrary points, they would comprise a set of ordered groups. The two sets of categories (one produced by the "other" model and one produced using the measure of need derived from the results of the state's AYP model) could be cross-tabulated and displayed as a table. Using a four-category accountability classification system as an example, the resulting table would be as follows.
 - Each school would fall into one of the cells in the table based on its assigned category on the "other" model (e.g., poor to excellent) and the measure of need (e.g., 1 for most in need of improvement to 4 for least in need of improvement) from the AYP model. A visual inspection of the numbers of schools falling into various areas of the table would provide information on agreement between the results of the two models.

FIGURE 7. MODEL FOR ANALYZING AGREEMENT BETWEEN TWO MODELS

		Category from "Other" Model			
		Poor	Fair	Good	Excellent
Category from AYP Measure of Need	(least in need of improvement)	4			
	3				
	2				
	1				
	(most in need of improvement)				

- Perfect agreement between the models would be reflected by values in cells A, B, D, and E and zeros in all the other cells. That pattern would, of course, be very unlikely.
- Good agreement would be indicated if there were values in the shaded cells and few if any values greater than zero in the unshaded cells.
- Poor agreement would be indicated if the values in the unshaded cells were greater than those in the shaded cells; the greater the values in cells G and H, the worse the agreement between the results of the two models.

Caution must be exercised regarding assumptions that the results of the actual run of the AYP model where schools are assigned to only two categories (met or not met) and are identified for improvement or more severe sanctions are "valid and reliable" based on relationships or cross-tabulations of data using one of the measures of need discussed in this section. An acceptable relationship using such a measure of need may be seen as an indication that AYP as specified in NCLB could be a valid and reliable model if it were amended to yield information that could better reflect the differing needs for improvement and allow appropriate sanctions to be applied based on that information.

Additional caveats apply when attempting to simplify methods for examining the validity of AYP models. While the NCLB AYP model establishes an overall AYP determination at the school level, there are some useful constructs that may be masked by the school level decision. Given a school that can be shown to be effective with the student aggregate (based on comparisons described earlier), that school may be "in need of improvement" regarding its effectiveness with certain subgroups. For example, further evidence about instructional practices in mathematics for students with disabilities (monitoring implementation of instructional strategies, direct observation), curriculum (IEP and standards alignment), and assessment (diagnostic evaluations) may suggest that the school

is in need of improvement at the subgroup level. The example above suggests that when each construct is evaluated separately, a body of evidence can exist to support two different conclusions about the school. The attempt to merge these two constructs into a single factor results in the introduction of error (false positive) as the subgroup construct amends the school classification.¹⁶

Stability and Reliability in the AYP Model

In this section, we first discussed ways for determining the level of need based on a state's AYP results and then ways of stating that need in terms of a set of categories or a numerical scale. That was followed by suggestions for ways that the measures of need could be used singly, or in conjunction with the results from other models, to judge the validity of the AYP model results. Next, ways of using the measures of need to examine the stability, or reliability, of the AYP model over time are discussed.

Readers may note that some of the information below is similar to the information presented in relation to AYP model validity. This is due to the substantive similarities between these two sets of issues. To ensure that each discussion remains complete while reading either section, these similarities have been retained.

There are no established procedures for calculating coefficients of reliability for AYP models. The procedures that follow provide some simple ways that the results from a state's AYP model can be examined to permit the state to estimate the reliability of its model. Throughout this section, reliability will be defined as the stability of AYP results over time (i.e., from year to year) and the terms "reliability" and "stability" will be considered to be synonymous.

The following discussions evolve from an examination of changes in a state's AYP results from year to year. It must be noted that some change (particularly, in a positive direction) is to be expected because all schools are working to improve student achievement. This may be reflected by overall changes in the distributions for the measures of need. Still, very large differences in the results from year to year would not be expected if the model is stable. Significant changes to the model itself or to data used in the model (e.g., changes in the statewide assessment program) will, of course, affect the degree to which year to year changes can be used as a measure of stability.

Probably the simplest way to use the measures of need for examining the reliability of the AYP model over time is to rank all the schools in the state on the selected measure. Taking only the schools falling at the top and bottom of the rankings (e.g., the top and bottom 10% or the top and bottom 25%), an informed decision could probably be made based on whether the schools in the high need and low need groups fall in the opposite group from one year to the next. Unless there are only slight differences in schools throughout the state, it is likely that there will be substantial differences in the needs measures of the schools in the high need and low need groups. If some of the schools in the high need group fall in the low need group the following year (and/or some of the schools in the low need group fall in the high need group), that would indicate a lack of stability over time.

Another relatively simple method for examining the stability of the AYP model is available if the state has collected reliable information concerning effective

¹⁶ Personal communication, J. P. Beaudoin, October 3, 2003.

schools and schools where there are problems affecting student achievement. Examples could include data from surveys or site visits. Using only the schools for which the effectiveness data are available, one could determine whether the majority of the identified schools fall in the correct group from year to year. If the identified schools fall in the correct group one year, but not the following year, that would indicate a lack of stability over time.

The stability of the AYP model can be estimated by comparing measures of need calculated for two consecutive years. Those comparisons can be conducted in various ways depending on the measures used:

- ♦ Using measures of need that can be assumed to comprise an interval or ordinal scale, some common statistical measures of relationship could be used. If the assumption of an interval scale is made, the relationship between the results of the AYP model over two years could be examined by running a Pearson product moment correlation analysis. If the correlation coefficient is moderate to high, the AYP model is producing stable results.
- ♦ Using measures of need comprising rankings of schools throughout the state, the relationship between the rankings produced by the AYP model over two years could be examined by running a rank correlation such as Spearman or Kendall.
- ♦ If the measures of need calculated from the results of the AYP model place schools into categories, a cross-tabulation procedure could be used. First, the measure of need would be used to place schools into different groups by dividing the school distributions from the two years into a number of categories. This could be done by using the quartile or quintile points in the school distribution or by simply dividing the score range of the measure of need into equal intervals (e.g., if values on the measure of need range from 10 through 90, “cut points” could be set at 30, 50, and 70 to form four groups of schools). While the categories formed using the measure of need would be based on somewhat arbitrary points, they would comprise a set of ordered groups. The two sets of categories (one from each year) could be cross-tabulated and displayed as a table. Using a four-category accountability classification system as an example, the resulting table would be as follows. A visual inspection of the numbers of schools falling into various areas of the table would provide information on agreement of the AYP results over time.

FIGURE 8. MODEL FOR ANALYZING AGREEMENT BETWEEN TWO YEARS

		Category from AYP Measure of Need Year 2			
		1	2	3	4
Category from AYP Measure of Need Year 1	4	G		F	E
	3			D	F
	2	C	B		
	1	A	C		H

- Perfect agreement between the AYP model results over time would be reflected by values in cells A, B, D, and E and zeros in all the other cells. That pattern would, of course, be very unlikely.
- Good agreement would be indicated if there were values in the shaded cells and few if any values greater than zero in the unshaded cells.
- Values greater than zero in cells G and H would indicate a lack of stability in the AYP model.

Since overall improvement in student achievement would be expected to occur from year to year, patterns with more schools in the higher categories in Year 2 might be expected even if the AYP model is stable. For example, some of the schools in category 1 during Year 2 might move into category 2 during the second year. There might be few schools in the top category during Year 1, but more schools in that category during Year 2.

The preceding discussion described some ways of examining the reliability of the AYP model in terms of the model's stability over time. Factors that can contribute to the reliability of the AYP model include the validity and reliability of the underlying student achievement measures. Validity of the assessments used in the AYP model is critical, but under an appropriate model for school accountability, adequate reliability of the model can be achieved with assessments with a "modest" level of reliability (Hill, 2001, p. 2).

Misclassification Error and Validity of an AYP Model

The earlier discussions illustrated procedures for looking at the overall (total school or LEA level) AYP determination and for using a measure of need to compare the results of the AYP model to the results from another accountability model or school rating system. Procedures for examining the stability and reliability of the AYP results over time were also discussed.

The discussion that follows addresses the causes of misclassification error in the AYP model and ways to decrease that error. The selection of a particular minimum “n” size for subgroups will not, by itself, ensure that the results of the AYP model are reliable (Kane & Staiger, 2002; Linn, Baker, & Herman, 2002; MacQuarrie, 2002). It is reliability, or consistency, of the AYP decisions that ED considers to be evidence of validity (ED, 2002). Given no specific procedure for controlling misclassification error (for individual NCLB subgroups or for schools), analyses can be conducted to determine the probable degree of error in the AYP results. That probable error rate could be compared to some criterion value to determine whether the results should be considered reliable (a necessary condition for the AYP model to be valid). It has been recommended that policymakers recognize, evaluate, and report the degree of uncertainty in accountability results (Baker & Linn, 2002; Linn, 2000).

It has been recommended that policymakers recognize, evaluate, and report the degree of uncertainty in accountability results (Baker & Linn, 2002; Linn, 2000).

There are many sources of error that affect the results of an accountability system. Many researchers agree that a chief contributor is sampling error (Cronbach et al., 1997; Hill, 2002; Hill & DePascale, 2002; Linn, 2001). The rationale is that the enrollment in a given grade or school in a particular year can be thought of as a sample of the total population of students who could attend, or could have attended, that school.

Some states are implementing procedures designed to reduce misclassification errors that are affected by sampling error. These procedures frequently involve the use of confidence intervals with or without a minimum n for group size. There are many ways that confidence intervals can be used within the AYP model and the most appropriate use depends largely on the characteristics of the AYP model implemented by the state. Although all NCLB compliant AYP models have to result in an overall AYP determination for each school and LEA, there are many differences in the ways that student assessment data and other variables are used to arrive at those determinations. A confidence interval is a range of values (e.g., percentage points) that bound an area within which a statistic (e.g., a school level percentage) might be expected to fall based on such factors as the number of data points (e.g., students) on which the statistic is based, and in the case of a proportion, how close to the target proportion the observed value falls. The appropriate application of confidence intervals can decrease the misclassification error within the AYP model.

Several researchers have examined the reliability of school ratings and accountability systems. While their studies set the context for examining the statistical reliability of AYP models, states will find it difficult to translate the findings into workable solutions for demonstrating reliability. Ligon, Jennings, and Clements (2002) demonstrate several methods for determining the reliability of AYP decisions for subgroups and discuss decisions regarding whether the results for certain subgroups are reliable enough to be reported and/or used within the AYP model. The authors acknowledge, however, that their work specifically addresses when to disaggregate students by the categories required in NCLB and does not address the determination of a school’s status as “in need of improvement” (i.e., the reliability of the school-level AYP determination).

Hoffman and Wise (2000) developed methods for determining the accuracy of student level classifications on the Kentucky Core Content Test. They employed the results of the score classification accuracy analysis to calculate standard

errors of measurement in the school-level numeric index used in the statewide accountability system. Standard errors of measurement varied by school size, value of the school-level numeric index, and within school variance of students. In the same study, the authors used a generalizability theory analysis as a second approach for establishing standard errors of measurement for each of several school sizes. Since separate standard errors were calculated for each assessment/grade/ subject combination, school-level standard errors could be estimated for any configuration of grades within a school.

Using the first approach, the authors calculated the probable number of points that each student would contribute to his/her school's numeric accountability index. After all students were processed, the result was the potential school-level score. An estimate of school level measurement error due to sampling error was calculated and a correction was applied for regression to the mean. The results from the calculations were estimates of error variances for grade and subject combinations at the school.

While Hoffman and Wise (2000) acknowledge that it is possible to use the above approach to calculate unique error variances for each school, they note that it may be too complicated to be practical. Using results produced by the above approach, they modeled the relationship between the school-level error estimates and three particular school characteristics: school size, mean observed student achievement level, and school variance in students' observed achievement. The school-level error composites were plotted against each of the three school characteristics in order to fit a set of polynomial regression equations. The authors concluded that, "a three-way classification of schools with multiple level category (e.g., large, medium, and small size schools crossed with high, medium, and low scores, crossed with high and low student variability) starts to become more complicated than simply reporting unique error estimates for each school." In either case, the result of the analyses for Kentucky was 18 separate standard errors of measurement, one for each of the grade/subject combinations on the assessment.

The second approach used by Hoffman and Wise (2000) was a generalizability analysis with school and years considered fixed and test forms and students assumed to be randomly sampled from an infinite domain. Analyses were conducted separately for three representative sizes of schools. The results of the analyses were three error variance estimates—one for each school size—for each of the 18 grade/subject combinations.

Whether derived from the student classification results or from generalizability theory, there existed SEMs (square roots of the error variances) only at the grade/subject level. The separate SEMs were combined (and weighted according to specifications in Kentucky statute) to estimate the error in total school accountability scores. The end result was standard errors of measurement for total accountability index scores for certain years. Assuming that errors are normally distributed, the SEMs can be converted to probabilities that estimate the likelihood that a school's true classification is within the assigned classification, is in a lower classification, or is in a higher classification, assuming that target scores (cut scores for delineating levels) are fixed without error.

The analyses conducted by Hoffman and Wise (2000) were specifically tailored to the assessment and accountability programs in Kentucky prior to the passage of NCLB. For a state to use their concepts for calculating misclassification errors

of its AYP model, the procedures would need to be adapted to fit the states specific assessment program and the logic used for making AYP determinations.

Hill and DePascale (2003) explicitly target the reliability of NCLB designs for adequate yearly progress. Their paper provides background on the requirements of NCLB, discusses sources of unreliability, contrasts the reliability of student-level vs. school-level results, and proposes several ways to determine decision consistency in an AYP model. Four methods are proposed for estimating the probability that schools will be correctly classified:

- ◆ Direct Computation – Calculate exact probabilities using areas under the normal curve and a set of standard equations.
- ◆ Split-Half – Randomly divide the students in the school into two groups and examine the consistency of the AYP decisions made on the two groups.
- ◆ Random Draws with Replacement – Draw random samples repeatedly from the students in the school, apply the AYP model to each sample, and examine the consistency of the AYP decisions.
- ◆ Monte Carlo – Use estimates of parameters for student assessment distributions, have a computer generate distributions, make repeated random draws, apply the AYP model to each sample, and examine the consistency of the AYP decisions.

Using the above methods (described in more detail in their 2002 paper) the authors estimate misclassification rates for schools of various sizes using student assessment instruments with various degrees of reliability and proficiency cut scores set at various points. The reliability of decisions made under compensatory and conjunctive rules is considered. Decisions regarding acceptable error rates and the tradeoff between Type I and Type II errors are discussed.

As with the paper discussed earlier, Hill and DePascale (2003) provide the conceptual detail required for understanding misclassification error and its importance with regard to the reliability of AYP models. The authors acknowledge that if the model of concern is complex (e.g., the NCLB AYP model that comprises many separate decisions across subject areas and subgroups), the direct computations for reliability and misclassification become too complex for actual use. The split-half method is subject to problems, particularly for small schools. While random draws with replacement is the best approach for complex models, that procedure may not be practical unless computer programs are set up to automate repeated sampling, run the AYP model on the samples, and capture the AYP results in a data file.

EXAMPLES USING ACTUAL AYP RESULTS

The following examples use the simple comparison procedures described earlier under the heading, Measures of Need and the Validity of AYP Models. The calculations are based on actual results from a state's AYP model and results from a separate model based on achievement and growth (AAG). The results are from the 2002-03 school year.

The overall school level results on the AYP model can be presented in terms of the number of schools that met AYP in all three areas—reading/language arts, mathematics, and other academic indicators. The reading/language arts and

mathematics AYP determinations were made using data from criterion-referenced tests administered to students in grades 3-8. Student achievement data were the percentages of students scoring at the proficient level or above on the state tests. The other academic indicator for high schools was graduation rate. For elementary and middle schools, the other academic indicator was a growth index based on the extent to which students in the school met growth expectations derived using multiple regression equations with previous year test scores as predictors.

The AAG (achievement and growth) model results can be presented in terms of the number of schools assigned each numeric rating (called a school performance classification). There are three possible levels: Levels 1 and 2 are considered below successful, Level 3 is successful, and Levels 4 and 5 are considered better than successful. As on the AYP model, the AAG model uses data from the criterion-referenced tests at grades 3-8. The student achievement data for the AYP model includes the percentage of students scoring at the basic level and higher as well as the percentage of students scoring at the proficient level and higher. A school growth composite is calculated based on growth expectations. The AAG model combines the achievement and growth components to yield a single numeric rating (or classification) for the school. Any school that meets its growth expectation is considered successful (is assigned at least Level 3). A more detailed description of the AAG model is not provided in this paper since it is not necessary for understanding the examples that follow.

The above AYP model results are presented in a two-level classification paradigm (met all three AYP variables or did not meet all three variables). The AYP results for each school can be compared to the school's results on the AAG model using a four-fold (2 by 2) table if the AAG results are collapsed to form a two-level paradigm¹⁷. Two different options seem reasonable. First, Levels 1 and 2 (below successful) could be combined and Levels 3-5 (successful and above) could be combined. Since the AYP model must be based only on students who score proficient or above on a state's tests (i.e., students scoring in the basic proficiency level do not count positively in the AYP model even if that level on the state's test represents adequate performance), the schools in Levels 4 and 5 could be combined to represent a higher level of school performance. The following tables show the cross-tabulations for model comparison results under the two paradigms described in this paragraph.

TABLE 12. COMPARISON OF THE AYP AND AAG MODELS UNDER PARADIGM 1

		AYP Determination (Results of the AYP Model)	
		Did Not Meet AYP	Met AYP
Results of the AAG Model	Successful (Levels 3-5)	283	396
	Below Successful (Levels 1 and 2)	129	13

Cell = number of schools
Total number of schools = 821

¹⁷ It is not necessary for the AYP results and the results from the other model to comprise the same number of levels or categories. A cross-tabulation table could be constructed regardless of the number of categories (e.g., "Met" vs. "Not Met" on the AYP model contrasted with the classification level on the AAG model using a 2 by 5 table containing 10 cells).

TABLE 13. COMPARISON OF THE AYP AND AAG MODELS UNDER PARADIGM 2

Cell = number of schools
Total number of schools = 821

Results of the AAG Model	AYP Determination (Results of the AYP Model)	
	Did Not Meet AYP	Met AYP
	High Performance (Levels 4 and 5)	Successful and below (Levels 1-3)
	100	271
	312	138

The data presented in Tables 10 and 11 show that, with one exception, many schools appear in each quadrant of each table. Agreement between the two models is indicated by schools falling in the shaded cells; disagreement is indicated by schools falling in the unshaded cells. The following calculations illustrate the level of agreement under the two paradigms.

TABLE 14. COMPARISON OF AGREEMENT UNDER PARADIGMS 1 AND 2

Total N = 821	Paradigm 1	AYP Determination (Results of the AYP Model)		Paradigm 2	AYP Determination (Results of the AYP Model)	
		Did Not Meet AYP	Met AYP		Did Not Meet AYP	Met AYP
Results of the AAG Model	Successful (Levels 3-5)	283	396	High Performance (Levels 4 and 5)	100	271
	Below Successful (Levels 1-2)	129	13	Successful and below (Levels 1-3)	312	138
Agreement	Number of schools with agreement: 396 + 129 = 525			Number of schools with agreement: 271 + 312 = 583		
	Rate of agreement = 525/821 = .64			Rate of agreement = 583/821 = .71		

Under Paradigm 1, the models agree concerning 529 (396 + 129) of the 821 schools. Therefore, the models agree about 64% of the time. Under Paradigm 2, the models agree concerning 583 of the 821 schools; this results in agreement about 71% of the time.

There was a high level of disagreement under both paradigms. While the overall agreement under Paradigm 2 was slightly higher, there was more consistency between the models in one cell under Paradigm 1 where almost none of the “below successful” schools met AYP.

Two additional examples are shown below. For both of those examples, the results of the AYP model were transformed into a “measure of need” as

described earlier in this section. The transformation procedure was carried out using the following steps:

1. Determine the number of subgroups (in reading/language arts and mathematics) for which schools were held accountable.
2. Delete schools from the analysis if there were too few subgroups to yield a reasonable range of percentage values (see Step 4).
3. Determine the number of subgroups at each school that met AYP. This determination considered the annual measurable achievement objectives, the testing participation rate, and, when appropriate, safe harbor calculations.
4. Divide the number of subgroups that made AYP by the number of subgroups for which the school was held accountable.
5. Divide the distribution of school level ratios calculated in Step 4 into a number of categories.
6. Assign each school to a category based on its calculated “met” ratio.

A total of 693 schools had between 6 and 12 subgroups for which they were held accountable for AYP. After calculating the “met” ratio for those schools, the resulting distribution was used to form two sets of categories—one set containing three categories and one set containing five. The categories were formed as follows:

FIGURE 9. EXAMPLE OF THREE- AND FIVE-CATEGORY GROUPINGS OF AYP RESULTS

Three-Category Scenario		Five-Category Scenario	
Category	Proportion of subgroups meeting AYP	Category	Proportion of subgroups meeting AYP
3	90 to 100	5	100
2	50 to 89	4	81 to 99
1	0 to 49	3	50 to 80
		2	10 to 49
		1	0 to 9

The following example compares the results of the AYP and AAG models using a three-category scenario. For the AAG model, Levels 1 and 2 were combined to form a “below successful” category, Level 3 comprised the “successful” category, and Levels 4 and 5 were combined to form a “high performance” category.

TABLE 15. COMPARISON OF THE AYP AND AAG MODELS USING THE THREE-CATEGORY SCENARIO

Cell = number of schools Total number of schools = 693		AYP Model Results (Category Based on "Need")		
		1 most need	2	3 least need
Results of the AAG Model	High Performance (Levels 4 and 5)	5	8	293
	Successful (Level 3)	8	23	232
	Below Successful (Levels 1 and 2)	38	40	46
Agreement		Exact (shaded cells only) = $(293 + 23 + 38) / 693 = 354 / 693 = .51$ (51%) Exact or adjacent (all cells except the upper left and lower right corners) = $(8 + 293 + 8 + 23 + 232 + 38 + 40) / 693 = 642 / 693 = .93$ (93%)		

The following example compares the results of the AYP and AAG models using a five-category scenario. For the AAG model, all school performance classification levels were reported separately.

TABLE 16. COMPARISON OF THE AYP AND AAG MODELS USING THE FIVE-CATEGORY SCENARIO

Cell = number of schools Total number of schools = 693		AYP Model Results (Category Based on "Need")				
		1 most need	2	3	4	5 least need
Results of the AAG Model (Level 1-5)	Level 5	0	2	2	2	116
	Level 4	3	0	4	6	171
	Level 3	5	7	14	14	223
	Level 2	5	29	14	6	43
	Level 1	14	9	1	2	1
Agreement		Exact (shaded cells only) = $(116 + 6 + 14 + 29 + 14) / 693 = 179 / 693 = .258$ (~26%) Exact or adjacent (all cells except the four uppermost left and four lowermost right) = $(2 + 116 + 4 + 6 + 171 + 7 + 14 + 14 + 5 + 29 + 14 + 14 + 9) / 693 = 405 / 693 = .584$ (~58%)				

In all the examples, the level of exact agreement between the two models ranged from 26% to 51%. As expected, higher levels of agreement were associated with comparisons involving smaller numbers of categories. There is no hard and fast rule for determining an adequate level of agreement. Each state would need to look at its results, consider the agreement ratios, and examine the tables to identify specific cells where the values seem reasonable or unreasonable.

However, while a high level of agreement indicates that both models are rating schools similarly, it does not necessarily mean that the ratings are correct or accurate; the ratings from both models could be incorrect. A low level of agreement indicates that the models are producing different results. In that case, one model may be producing correct results or both models may be producing incorrect results.

The most compelling evidence regarding the accuracy of the results from an AYP model would be high agreement between those results and some kind of verified measures of school effectiveness. If such measures are available, the procedures above could be applied to the school effectiveness measures the way they were applied to the AAG model results. AYP results were available for only one year. Had two years of AYP results been available, the examples above could have been applied for examining the stability of the model over time.

SUMMARY AND CONCLUSIONS

While several studies have led the way toward an understanding of the reliability and validity of accountability systems, there is no general paradigm that can easily be applied to states' AYP models to demonstrate they are "reliable and valid." Some simple methods were outlined in this section that states can use for examining the relationship between their AYP results and other trusted measures of school performance. Those same methods can be employed for examining the stability of the state's AYP results from year to year.

REVIEW AND APPEALS PROCESSES

In addition to considering the reliability and accuracy of the accountability decisions as described above, states must provide schools and LEAs with an opportunity to formally dispute the outcomes of the decisions. This opportunity not only allows for the consideration of data that cannot be part of the AYP model but also gives the entities that are directly affected by the AYP results a voice in the system. Thus, use of an appeals process can enhance the validity, fairness, and credibility of the system.

An appeals process can encompass two categories of review, including—

- ♦ reviews and corrections of raw data and indicators prior to the AYP analyses
- ♦ reviews and appeals of preliminary AYP results

In addition, states should create and disseminate a calendar that specifies the dates by which—

- ♦ data are to be submitted to the state
- ♦ data and/or indicators will be available for LEA review
- ♦ any corrections to data and indicators must be completed

- ♦ data and indicators will be considered final
- ♦ preliminary AYP results will be available
- ♦ appeals must be filed
- ♦ state responses to appeals will be issued
- ♦ AYP results will be considered final.

REVIEWS AND CORRECTIONS OF RAW DATA AND INDICATORS

The reviews of raw data and indicators relate to the first four of the calendar points noted above, when—

- ♦ data are to be submitted to the state
- ♦ data and/or indicators will be available for LEA review
- ♦ any corrections to data and indicators must be completed
- ♦ data and indicators will be considered final.

As described in the section on indicators, states must verify that the data used to create accountability indicators are accurate. In states that collect raw data from LEAs and then generate the accountability indicators, this means that states should create summary statistics (e.g., total and disaggregated enrollment counts) and at least some of the accountability indicators (e.g., percent tested, percent proficient, attendance rate) for LEAs to review prior to conducting the AYP analyses. In states where LEAs enter data into the database used to generate these statistics, states should impose and enforce a deadline by which these data are to be entered and considered final, in addition to providing accountability indicators for LEAs to review.

States vary in their policy positions on changing data after the deadlines have passed. Certainly, no one would like to use data that are known to be inaccurate when making a high-stakes accountability decision. However, if states remain open to data changes past a necessary deadline, they will find it very difficult to complete the required AYP calculations. In addition, states that are strict about their deadlines may experience fewer problems in getting their data on-time in the future.

REVIEWS AND APPEALS OF PRELIMINARY AYP RESULTS

The reviews of raw data and indicators relate to the last four of the calendar points noted above, when—

- ♦ preliminary AYP results will be available
- ♦ appeals must be filed
- ♦ state responses to appeals will be issued
- ♦ AYP results will be considered final.

Following this timeline, it should be clear that any AYP result that has not been reviewed by the LEA and school in question should be considered preliminary.

States vary widely in how they handle the appeals process. Of the states that have well-developed processes in place, notable characteristics are apparent:

- ♦ **Clear deadlines for filing appeals**

- ♦ **Readily available forms**

At least one state currently provides an on-line form that all schools and LEAs must complete and submit as part of the AYP decision process. If the school or LEA is not appealing its identification, the principal/superintendent must indicate that the AYP results are accepted and sign-off on the form.

- ♦ **Limited allowable reasons for appeal**

It is wise for states to consider what should and should not be grounds for appeal in their contexts. For example, some states allow all “small” schools that are identified for improvement an automatic appeal without requiring any other justification. Anticipating concerns about results for SWDs and LEP students, some states will allow appeals for any school that is serving a relatively high proportion of SWDs or LEP students with documented lower levels of English proficiency.

Some states that use well-established, statewide student-level information systems will not allow appeals on the basis of faulty data that LEAs have already reviewed and approved. On the other hand, a state may wish to allow appeals of AYP results in the event of an unforeseeable situation, such as a documented loss of test materials in the shipping process, a weather event, or other clear emergency that seriously affected the testing process. One state that allows such appeals specifically denies appeals in cases where school or LEA staff have knowingly violated test security.

- ♦ **Specific requirements for the types of data that will be considered in the state’s review of the appeal**

If LEAs have already approved the data and indicators, then the major remaining issue for appeals is student achievement. States should specify the types of alternative data that schools and LEAs can submit as evidence to support their appeals. For example, a state may allow results from locally-administered norm-referenced tests but not results from a locally-developed assessment of unknown quality. In addition, the state should indicate what level or increase in performance would be sufficient to warrant a possible change in identification status, particularly since locally-administered norm-referenced tests are not likely to be fully aligned with the standards as well as results from the state assessment will be.

SECTION 4

EVALUATING THE CONSEQUENCES

Purpose

To evaluate how the applied consequences associated with accountability decisions are implemented and related to both intended and unintended emergent changes in school and LEA functioning.

Major questions

- ♦ How well are rewards, sanctions, and interventions implemented?
- ♦ How do school and LEA characteristics, as well as other facets of the context, moderate the implementation of the consequences?
- ♦ To what degree are the intended actions occurring in relation to the application of rewards, sanctions, and interventions?
- ♦ To what degree are negative, unintended consequences occurring in relation to the application of rewards, sanctions, and interventions?
- ♦ To what degree are the reform activities associated with achievement of the goals of the system?

"No matter how accurately and consistently the accountability system is able to identify failing schools, it just becomes an academic exercise if nothing happens to improve the work of schools."

(Marion & Gong, 2003)

CONSEQUENCES IN ACCOUNTABILITY SYSTEMS

The accountability system has been designed, the indicators reviewed, and the decisions evaluated. Now what?

As tempting as it might be to stand back and cross one's fingers in the hopes that everything falls into place and leads to improved overall achievement, closed achievement gaps, and other noble goals, it is at this point that the most challenging and critical validation work must begin. It simply cannot be assumed that selecting the "right schools" and assigning them pre-specified consequences will lead to the intended reforms and ultimately result in the achievement of intended goals. Nor can it be assumed that it will not lead to any unintended, negative consequences. As indicated in the outset of this paper, these are hypotheses (part of the Theory of Action) that must be tested.

The primary concerns with regard to accountability consequences are—

- ♦ how the consequences that are imposed on schools, which include rewards, sanctions, and interventions, are implemented;
- ♦ whether imposed consequences are associated with the emergence of the intended reforms, as indicated in the state's Theory of Action, and also the emergence of any negative unintended conditions or activities; and
- ♦ whether the activities and conditions that emerge after the application of consequences are associated with the achievement of the accountability goals.

To address these concerns will require—

- ♦ a careful mapping of this portion of the Theory of Action, including the conceptual and operational specifications of the—
 - consequences that are imposed on schools and LEAs as part of the accountability system;

- activities that are intended to be associated with the application of consequences; and
 - negative, unintended but plausible conditions or activities that could be associated with the application of consequences.
- ♦ use of multiple, coordinated data collection strategies over several years.

Table 17 provides an example of an initial mapping of the Theory of Action for the consequences component of an accountability system. This information could then be used to design a set of studies to determine whether the imposed consequences were being implemented as intended, whether these consequences were associated with the intended reforms, and whether any negative, unintended consequences emerged.

Following this table are a series of sample studies that illustrate how states could approach parts of the validation process for accountability consequences. Each of these cases focuses on a different issue of concern:

- ♦ Case A focuses on **compliance** in sanction implementation
- ♦ Case B looks into **changes in curricular alignment** that occur after identification for improvement

Case C addresses **professional development** after identification for improvement

These issues have been separated for the sake of clarity here, but states are probably interested in combinations of these and other issues and could combine study elements to address multiple issues in any one study. For each case, some background information is provided, the state's primary focus is identified, and the state's approach is described.

TABLE 17. SAMPLE LISTING OF IMPOSED, EMERGENT, AND PLAUSIBLE NEGATIVE CONSEQUENCES

Imposed consequence and its intended implementation	Conceptual description of how this is supposed to work	Operational description of how this is supposed to work (Theory of Action and Emergent Consequences)	Potential implementation problems and negative consequences
Reward Monetary award of \$10,000 for 10 schools making the greatest achievement gains. Award to be distributed by October 1 and can be used in any way the school chooses.	The possibility of receiving the award should motivate school faculty to institute reforms to improve achievement.	Eligible schools should be awarded their money prior to the start of the following school year. When school faculty become aware of the potential award, they should— - improve the alignment of their curricula with the standards; - make greater use of instructional strategies that have been shown to be effective; - engage in activities that increase students' motivation to perform well on the tests	- Eligible schools do not receive the money or receive it late. - Driven by the desire to obtain the award, school faculty could use inappropriate test preparation and/or administration techniques. - If faculty don't think the school has a chance to succeed, they may not work hard to institute reforms. - More students with disabilities or limited proficiency in English may begin taking alternate assessments.
Sanctions Schools not making AYP for 2 consecutive years must offer students choice and LEAs must pay for necessary transportation. Schools not making AYP for 3 consecutive years must offer students choice and provide supplemental services for students.	The threat of the sanction should motivate school faculty and LEA personnel to institute reforms to improve achievement. The actual imposition of the sanction should allow students to receive better instructional services than are available in their original schools.	LEAs and schools should— - make sure staffing, other resources, space, and transportation necessary for choice are in place - make sure supplemental service providers and the resources necessary to pay them are available - improve the alignment of their fiscal resources with their reform strategies - improve the quality of professional development activities and align these activities with reform strategies - raise the minimum requirements for paraprofessionals - improve the alignment of their curricula with the standards - make greater use of effective instructional strategies - improve the coherence of their instructional programs - engage in activities that increase students' motivation to perform well on the tests	- All schools within a reasonable travel zone will be under this sanction so the choice is between two improvement schools. - No school in a reasonable travel zone has room for additional students or surrounding LEAs refuse to accept choice students. - Service providers may not be available in all areas or have the capacity to assist all eligible students. - The cost of providing supplemental services may exceed the available funds so some eligible students may not receive services. - Driven by the desire to avoid the sanction, school faculty could use inappropriate test preparation and/or administration techniques. - Paraprofessionals may be in short supply when employment requirements increase but funding for increased salaries does not. - More SWDs or LEP students may begin taking alternate assessments. - Programs and courses not seen as critical to reading or math test performance may be eliminated. - Practices for coding student data may be modified such that the number or type of students whose scores count in AYP is changed inappropriately.
Intervention For each school identified for improvement, the state provides half-time master teacher in reading and in mathematics as well as support from a support team.	The additional support should help troubled schools to improve the quality of their instruction, which should lead to improved achievement.	The state identifies and pays master teachers for each school. These master teachers improve the quality of instruction in the school by— - helping to review and revise curricula - facilitating coordination across grades and classrooms - providing on-site professional development support that is integrated with the LEA PD program - acting as mentors to newer teachers	- The state does not have the resources to support master teachers for the large number of schools identified for improvement. - Not enough individuals apply or qualify for the master teacher positions. - Some identified schools are already implementing commercial school reform designs and master teachers may not be familiar with or in agreement with these designs. - The needs in some schools are so great that a half-time position is not enough to be effective.

EXAMPLES OF CONSEQUENTIAL VALIDATION STUDIES

CASE A: *WERE THE SANCTIONS ACTUALLY IMPOSED AND WHO EXERCISES THE OPTION OF SCHOOL CHOICE?*

Background

State A did not have a formal, statewide accountability system in place prior to the 2000-01 school year. Accountability previously had been limited to Title I schools only and addressed process-related compliance issues. As a result, neither the state nor its LEAs have any solid experience with the imposition of sanctions related to outcomes, let alone in the evaluation of reform activities at the school or LEA level. At the time of the study (the 2003-04 school year), 96 schools were identified for improvement; these were distributed across 24 LEAs. Seventy-nine of the schools were in their first year of improvement, ten were in their second year, six in their third year, and one in its fourth year at the time of the study.

Design

As a first step in evaluating the consequences of its accountability system, State A decided to develop a plan for determining whether the sanctions associated with failure to make AYP for multiple years were actually being implemented. The state conceptualized this as the first of three levels of study. Level two would involve the degree to which schools and LEAs engaged in the intended reform activities once the sanctions were imposed, and level three would address the relationship between engagement in reform activities and improvement in student achievement. The state planned to begin studies related to levels two and three in the following school year. The state was still working on its Theory of Action and needed time to specify what and how reforms would contribute to improved achievement. Thus, the delay would allow time for study design and for adjustments in the work load of some key staff.

In addition, there were relatively strong political movements supporting school choice and charter school options in the state. The State Board of Education (SBE) asked the state's evaluation office to look into the issue of choice and how it works in practice. The SBE was interested in finding out how, when, and why families take advantage of choice options and how statewide choice might affect enrollments in different areas.

To address their level one concern about simple compliance as well as some of the SBE's questions related to school choice (the first sanction for Title I schools in improvement), the state identified the following overarching questions for a one-year study:

- ◆ How and when are schools and LEAs notified of their sanction(s)?
- ◆ How and when do schools and LEAs prepare for the implementation of a sanction?
- ◆ How and when do sanctions actually take effect?
- ◆ What factors affect sanction implementation?

Since the answers to these questions would likely vary by the type of sanction, the state further refined these questions and the information necessary to answer

them by sanction type, as illustrated in Table 16. This would allow the state to look separately at choice, though only in those schools required to offer it due to low performance, as requested by the SBE.

The state determined that it would extensively study the compliance issue for only one year. The insights gained into the implementation of the sanctions would then contribute to improving support for implementation and to the development of a relatively simple form that all improvement schools would complete and submit as evidence of their sanction implementation. The state's resources for the study were limited. Two staff members were assigned to conduct the study and decided to obtain documents and input from the LEAs only. No surveys or on-site data collections would be involved due to budget constraints, although the state is considering conducting some focus groups next year to help them understand why students who are eligible for choice or supplemental services do or do not take advantage of these opportunities. The state obtained assistance with data collection, analysis, and interpretation through a contract with a state university research center.

TABLE 18. SAMPLE QUESTIONS AND EVIDENCE FOR CASE A

Sanction	Conceptual Description	Specific Evaluation Questions	Sample Evidence
Year 1 School choice with LEA providing transportation (79 Schools)	Students in schools identified for improvement will be offered the opportunity to attend another school that may provide better instructional services.	- How and when are schools and LEAs notified that choice must offered? - How and when do schools and LEAs prepare for the implementation of a sanction?	State's notification process and timeline, including evidence of LEA receipt - Documentation indicating how many students could be taken at each alternate school by grade level and of contacts with surrounding LEAs for taking students as necessary - Copies of parental notification letters - Documentation of LEA plans for funding transportation necessary to implement choice Report indicating how many students took advantage of choice, where they opted to attend, and when they enrolled - Demographic make-up of the schools and LEA - Demographic characteristics of students who choose to change schools - State laws and court mandates related to desegregation - Proportion of identified schools in the LEA - Proximity of identified schools to one another - Proximity of identified schools to non-identified schools in the LEA - Proximity of non-identified schools in neighboring LEAs - Academic programs in place in the school - Availability of non-academic family resources in the school (e.g., health clinic) - Participation in parent/ family organizations (e.g., PTA)
Year 2: School choice plus supplemental services (10 schools)	Students in schools identified for improvement will be offered the opportunity to attend another school that may provide better instructional services. Remaining low-income students will be offered appropriate supplemental services by state-approved supplemental service providers (SSPs).	- How and when are schools and LEAs notified that choice must offered? - How and when do schools and LEAs prepare for the implementation of a sanction? - How and when do sanctions actually take effect? - What factors affect sanction implementation?	State's notification process and timeline, including evidence of LEA receipt All those listed for Year 1, plus - List of SSPs available to provide services in each identified school - Report of how many students were eligible for services, the estimated costs, and the available resources to cover these costs Report of how many students received services and the initiation, frequency, and duration of these services Above reports on— - eligibility and cost - SSP availability

TABLE 18. SAMPLE QUESTIONS AND EVIDENCE FOR CASE A (CONTINUED)

Sanction	Conceptual Description	Specific Evaluation Questions	Sample Evidence
Year 3: All the above plus corrective actions (6 schools)	Students in schools identified for improvement will be offered the opportunity to attend another school that may provide better instructional services. Remaining low-income students will be offered appropriate supplemental services by state-approved providers. At the same time, the school must engage in corrective actions, such as staff replacement or implementation of new curricula.	How and when are schools and LEAs notified that choice must be offered?	State's notification process and timeline, including evidence of LEA receipt
		How and when do schools and LEAs prepare for the implementation of a sanction?	All those listed for Years 1 and 2, plus - State requirements for corrective actions under different circumstances - Documentation of LEA plans for identifying and initiating corrective actions
		How and when do sanctions actually take effect?	LEA implementation calendar for each corrective action
		What factors affect sanction implementation?	- Union protection of staff - Resources and guidance available to support curriculum refinement
Year 4: Restructuring (1 school)	The school must "initiate plans for restructuring" which may include "opening the school as a charter ... replacing all or most of the ... staff...or turning over school operations either to the state or to a private company with a demonstrated record of effectiveness."	How and when are schools and LEAs notified that choice must be offered?	State's notification process and timeline, including evidence of LEA receipt
		How and when do schools and LEAs prepare for the implementation of a sanction?	All those listed for Years 1, 2, and 3, plus - State requirements for restructuring - Documentation of LEA plans for restructuring
		How and when do sanctions actually take effect?	State and LEA implementation calendar for restructuring
		What factors affect sanction implementation?	- State plans, resources, and requirements for restructuring - State laws and requirements for charter schools - Union protection of staff - Availability of private companies for single-school takeovers

CASE B: *HOW IS CURRICULAR ALIGNMENT RELATED TO PERFORMANCE AND IMPROVEMENT STATUS?*

Background

State B has had a statewide accountability system in place for several years. However, prior to NCLB, the state had only identified a total of 37 schools for improvement in seven years; 15 of these had already transitioned out of improvement status prior to the 2002-03 school year. The state had intentionally designed its old AYP model to identify only a small number of schools because the state wanted to provide intensive, team-centered support to the identified schools and their LEAs and did not have the staff or other resources to do so with a large number of schools. Now, after the 2002-03 school year and the application of the NCLB AYP model, 287 schools were identified; all but one LEA had at least one school in improvement status. The state uses the NCLB AYP model as one component of its accountability decision process for schools. Schools that meet their AYP targets and also meet other criteria can earn a performance rating of 4 or 5 in a 5-level rating scheme.

Design

The state could no longer provide the intensive support in the same style it once had, but wanted to retain a strong focus on improving classroom practice. In fact, in its Theory of Action, improvement of classroom practices—meaning better alignment of curricula with standards and enhancing teachers’ repertoires of effective instructional strategies—is the primary catalyst for improving achievement. To ensure that the neediest schools would still get intense attention on classroom practices, the state developed a triage system to categorize schools by type and degree of need and assigned “critical care” teams to the neediest schools in each content area.

The state also adopted a study plan that would provide helpful information at the state, LEA, and school levels and distribute responsibility for support and data collection across these levels. Since the state was already preparing to study alignment between its voluntary state standards and assessments, state staff decided to expand this study to include a measure of the enacted curricula. That is, the state wanted to know, and wanted its educators to know, how the curricula used in the classrooms compared with the standards and with the assessments. By using the existing measures offered as part of CCSSO’s Surveys of Enacted Curriculum (SEC) program that the state was already participating in, state staff were able to begin gathering data immediately.

The state identified the following questions to guide their study:

- ♦ How well aligned are the intended and enacted curricula, by content area?
- ♦ How well aligned are the intended curricula and the state assessments, by content area?
- ♦ How well aligned are the enacted curricula and the state assessments, by content area?
- ♦ Does the quality of alignment vary across content areas (reading and mathematics)?

- ♦ How does alignment differ across school performance levels and triage categories?
- ♦ Does the quality of alignment vary across grades?
- ♦ What other factors affect alignment between intended and enacted curricula?
- ♦ Does the quality of alignment vary over time?
- ♦ What factors support the improvement of alignment over time?

Three types of data collection were involved in this study. First, the state engaged in a process for coding and analyzing the content of the statewide curriculum frameworks and statewide assessments in reading/language arts and mathematics. Money from the assessment budget was used for these costs, since the alignment issue was a major focus in the assessment validation process. Second, surveys were administered to teachers in all participating schools. The third type of data collection involved the use of extant data from state and LEA databases. To gather evidence in relation to whether alignment changes over time, the state committed to collecting survey information in the spring of each school year for the next five years. Non-identified schools could opt for participation every other year.

Participation in the study was required for schools in improvement—although voluntary and anonymous for individual teachers—and also open on a voluntary basis to all other schools and LEAs. The state hoped that a number of non-identified schools would participate so that comparisons could be made between identified and non-identified schools. The state encouraged LEAs to participate and to get their schools and faculty involved by covering all costs associated with the study, highlighting the rich information that schools would receive by participating, offering incentives to participating schools for improvement in achievement, and implementing a protocol for data collection that ensured teacher anonymity.

The state assigned three staff to the study—one each from the English/ language arts and mathematics curriculum teams and one from the evaluation unit. The state contracted with the SEC program and with its own regional lab for assistance with the study.

The state paid for the surveys and offered training for up to three personnel from each LEA in the administration, interpretation, and use of the survey and other study data. This training was presented using a trainer-of-trainer model so that the LEA staff could provide in-turn training to school team leaders. Training took place in two parts. First, LEA participants were brought together to discuss the purpose of the study, the anticipated benefits of participation, and the administration of the surveys. The state solicited feedback during this meeting that was used to refine the data collection process (e.g., some LEAs wanted both paper-and-pencil and on-line response options for the surveys, but others only wanted the on-line version). Second, after the coding and surveys were complete, the LEA trainers were reconvened to learn how to interpret the results and begin thinking about how to use the results to improve instruction. By providing this kind of training and support, the state integrated support for school reform with one study in the evaluation of its reform model.

TABLE 19. SAMPLE STUDY PLAN FOR CASE B

Question	Hypotheses	Data required (separate analyses for R/LA and mathematics)	Description of analyses (separate analyses for R/LA and mathematics)
How well aligned are the intended and enacted curriculum , by content area?	The intended and enacted curricula will be moderately to poorly aligned.	• Coded state frameworks • Survey results	At the state, LEA, and school levels, compare the patterns of emphases between the standards and the curriculum teachers report using in their classrooms.
How well aligned are the intended curriculum and the state assessments , by content area?	The assessments will be moderately to well aligned with the intended curricula (state standards).	• Coded state frameworks • Coded state assessments	At the state level, compare the patterns of emphases between the standards and the statewide assessments.
How well aligned are the enacted curriculum and the state assessments , by content area?	The enacted curriculum and assessments will be moderately to poorly aligned.	• Survey results • Coded state assessments	At the state and LEA levels, compare the patterns of emphases between the assessments and the curriculum that teachers report using in their classrooms.
How does alignment between intended and enacted curricula differ across school performance levels and triage categories ?	Alignment will be better at the higher performance categories and lowest in the most critical triage category.	• Coded state frameworks • Survey results • School performance level • School triage category (if the school was identified for improvement)	Compare alignment indices across performance levels and across triage categories.
Does the quality of alignment vary across reading and mathematics content areas?	Alignment will be comparable across content areas.	• Coded state frameworks • Coded state assessments • State survey results	Compare alignment indices across content areas.
Does the quality of alignment vary across grades ?	Alignment will be better in the lower grades than in the higher grades.	• Coded state frameworks by grade level by content area • State survey results disaggregated by grade level	Compare alignment indices across grades.
What other factors affect alignment between intended and enacted curriculum?	Alignment will be better where there are across-grade curriculum teams, content-focused professional development, and lower levels of staff turnover.	• Coded state frameworks • Survey results • Description of LEA approach to professional development (from addendum to survey) • Indicator of use of across-grade curriculum teams (from addendum to survey) • Staffing data related to turnover	Across LEAs and schools, compare alignment indices while accounting for use of across-grade curriculum teams, content-focused professional development, and lower levels of staff turnover.
Does the quality of alignment vary over time ?	Alignment will improve over time.	• Coded state frameworks • Survey results for each year	Compare alignment indices across years.
What factors support the improvement of alignment over time?	Alignment will improve where there are across-grade curriculum teams, content-focused professional development, and lower levels of staff turnover.	• Coded state frameworks • Survey results for each year • School performance level • School triage category (if the school was identified for improvement) • Description of LEA approach to professional development (from addendum to survey) • Indicator of use of across-grade curriculum-teams (from addendum to survey) • Staffing data related to turnover	Across LEAs and schools, compare alignment indices over time while accounting for school performance level, triage category, use of across-grade curriculum teams, content-focused professional development, and lower levels of staff turnover.

CASE C: *HOW IS TEACHER QUALITY AND THE QUALITY OF PROFESSIONAL DEVELOPMENT RELATED TO ACCOUNTABILITY PERFORMANCE?*

Background

State C had a relatively rigorous accountability system in place prior to NCLB. However, the NCLB AYP model resulted in the identification of about 20% more schools than its previous decision model. The state's existing accountability policies focused on support rather than sanctions, so the sanctions-oriented consequences component of their accountability system would be new. The process for working with schools in need of improvement was already rather institutionalized and was very much aligned with long-held values in the state. After much discussion, the state decided not to radically change its improvement and support system with the implementation of NCLB even though the number of schools to be served would continue to increase over the next several years. Instead, the state expanded its collaboration with the regional agencies across the state and its existing relationship with a university collaborative.

Design

The state would be focusing attention in three areas of study this year. First, the state would be conducting several studies of the AYP model, which was very different from its previous process for identifying schools. For example, it would be comparing NCLB results with results from its previous model, which it would continue to "run in the background." Second, the state would conduct a series of data audits every year, and each LEA would continue to be audited at least once every five years. Third, the state decided to address, in turn, the factors it believed contributed to improved student achievement as specified in its Theory of Action.

State C had invested heavily in improving the quality of its teaching force. It had revised certification requirements, raised the pay scale, implemented an extensive system of new teacher mentoring and evaluation, and required LEAs to use research-based strategies for improving the quality and effectiveness of their professional development (PD) programs.

With regard to its accountability system, teacher quality and professional development figured prominently in State C's Theory of Action. State education leaders considered these to be critical to their standards-based reform efforts which ultimately were supposed to lead to high student achievement. Better prepared and supported teachers would be better able to teach to the standards (Smith & O'Day, 1991). The state was interested in testing this relationship but was also very concerned that the rapidly escalating AYP targets and the new emphasis on sanctions could negatively affect the progress made in supporting and raising expectations for teachers. In schools where the targets were already challenging, teachers could become disheartened if their increasing capacity cannot translate quickly enough to keep pace with increasing expectations for student performance. It may become more difficult for such schools to attract and keep experienced teachers as the possibility of restructuring nears. And, although being identified for improvement under the previous accountability system was not a reason for celebration, it had not had the negative connotations now associated with identification.

The state decided to conduct a study to examine these issues and formulated the following questions:

- ♦ Is the quality of PD programs associated with higher student achievement?
- ♦ Is the proportion of schools identified for improvement associated with changes in the quality of PD programming?
- ♦ Do teachers' motivation and attitudes change with identification status?
- ♦ Does recruitment change with identification status?
- ♦ Does retention change with identification status?

The state conceptualized a five-year study to address these questions and decided to use a combination of case studies, surveys, and reviews of extant data. Six LEAs were selected and agreed to participate as cases. In each of these LEAs, members of the state study team reviewed the PD plan and observed three different types of PD activities throughout the year. All the teachers and administrators in these LEAs were asked to complete short surveys twice each year. Extant LEA recruitment and retention records were also reviewed and existing files for student achievement were retrieved for participating LEAs. The state's study plan is illustrated in Table 18.

The state used a research-based approach to evaluating the quality of the PD programs. Using the PD plan reviews, observations, and survey data, each of the LEA's PD programs was rated along the dimensions of form, duration, collective participation, active learning, coherence, and content focus (Garet, Porter, Desimone, Birman, & Yoon, 2001).

TABLE 20. SAMPLE STUDY PLAN FOR CASE C

Question	Hypotheses	Data Collection
Is the quality of professional development programs associated with higher student achievement ? (LEA-level analyses)	LEAs with higher quality professional development programs would have higher levels of student achievement.	• PD plan • Three PD observations • Teacher and administrator surveys • Student performance on reading and mathematics state assessments
Is the proportion of schools identified for improvement associated with changes in the quality of professional development programming? (LEA-level analyses)	The quality of PD will decrease as the proportion of identified schools increases because of the greater burden on state and LEA resources.	• PD quality ratings (developed using data sources listed above) • Proportion of schools identified for improvement
Do teachers' motivation and attitudes change with identification status? (School-level analyses)	Teachers' motivation and attitudes will diminish as the number of years in identification increases. They will feel less interested in actively participating in PD and using the content and strategies learned in PD in their classrooms.	• School-level identification status • Teacher survey responses to motivation and attitude items
Does recruitment change with identification status? (LEA- and school-level analyses)	LEAs will find it more difficult to attract highly qualified teachers to schools as the proportion of schools in identification increases.	• School-level identification status and LEA-level proportion of schools identified • LEA recruitment and hiring records
Does retention change with identification status? (LEA- and school-level analyses)	Teacher retention will decrease as the number of years in identification increases.	• School-level identification status and LEA-level proportion of schools identified • LEA retention records

CONCLUSIONS ABOUT THE STUDY OF CONSEQUENCES

These three short cases focused on a few of the infinite possibilities for studying the complex chain linking AYP decisions to improvements in student achievement. Some key ideas can be drawn from these cases:

- ♦ Questions can range from the very basic (e.g., “How and when are schools and LEAs notified that choice must be offered?”) to the rather sophisticated (e.g., “What factors support the improvement of alignment over time?”).
- ♦ Studies of consequences can piggyback on existing reform activities, evaluations, data collections, and auditing mechanisms.
- ♦ States can prioritize their evaluation agenda by considering such issues as availability of resources and the need for evidence to inform policy decisions.
- ♦ States should develop or take advantage of relationships with universities, within-state regional education agencies, regional labs, and outside consultants to get help with the aspects of some studies.

Summary

The information provided in this paper is meant to assist state and local educators in conceptualizing their plans for examining validity in their accountability systems. It is a framework only. The details of any validation plan will depend upon the design of the system under investigation as well as the perspectives and resources of the agencies conducting the evaluation.

Major points from this paper are summarized in this section. A more comprehensive summary is provided in the *Executive Summary* that is published separately.

BACKGROUND

States and LEAs across the country are now implementing accountability systems that must meet a number of highly specific requirements under NCLB. States have an obligation to evaluate these systems to determine whether they are associated with achievement of the intended goals and not with unintended, negative outcomes. This obligation is especially critical because these systems encompass high stakes consequences for schools and LEAs that do not meet certain criteria. In addition, the tight timeline under which these requirements must take effect has meant little time for planning and preparation prior to implementation. This combination of high stakes and limited planning and preparation time means that errors may be both more likely and more costly.

One of the first questions that policymakers and staff may pose with regard to accountability systems is whether the “right schools” are being identified for improvement. However, the answer to this question depends on the answers to several other questions:

- ♦ What kinds of schools were meant to be identified?
- ♦ What are the goals the accountability system is intended to achieve?

- ♦ How trustworthy are the data on which the decision was based?
- ♦ How were these data combined in the decision-making model? What happens once schools are identified?
- ♦ What consequences and reforms are implemented and are they appropriate and effective?

To address these questions will require the accumulation and evaluation of a large body of evidence. When carried out systematically and rigorously, this process can be considered **validation**.

FOUNDATIONS

Accountability systems can be defined in the following way:

Accountability systems are used to achieve specific educational goals by attaching to performance indicators certain consequences meant to effect change in specific areas of functioning.

This definition encompasses four critical concepts:

1. Performance indicators
2. Decision rules for attaching consequences to performance indicators
3. Consequences (imposed and emergent)
4. Goals

A Theory of Action specifies how these components are interrelated and how the entire accountability system is meant to work. A validation process must involve studies of each of the parts and of how they relate to one another.

WHY VALIDATE ACCOUNTABILITY SYSTEMS?

When a state implements an accountability system, it is essentially making a claim that the system will help to achieve specific educational goals (e.g., improved achievement in reading/language arts and mathematics). Unless the state evaluates its system, it can have no evidence to support this claim. The state would not be able to defend its accountability decisions or the imposition of sanctions and would almost certainly lose credibility with its stakeholders and challenges in court. It is impossible to tell why a system does or does not work—or to defend the system against perfectly reasonable and inevitable challenges from affected stakeholders—unless it has been evaluated systematically.

WHAT IS VALIDITY?

Validity with regard to accountability systems refers to the degree to which evidence and theory support the indicators, decisions, and consequences, individually and combined as established via the theory of action, as used for the purpose of achieving specific goals. “An accountability system can be said to have validity when the evidence is judged to be strong enough to support the inferences that: [1] the components of the system are aligned to the purposes, and are working in harmony to help the system accomplish those purposes; and [2]

the system is accomplishing what was intended (and did not accomplish what was not intended)” (Marion et al., 2002, p. 105).

Accountability is not Adequate Yearly Progress (AYP). The overarching validity question is not “Does this accountability system pick the right schools?”, but, rather, “Does this accountability system do what it is intended to do?” Selecting the right schools is only part of the answer.

A VALIDATION FRAMEWORK

The basic elements of accountability systems (performance indicators, decision rules, consequences, and goals) are related to one another according to a theory of action. The validation framework presented in this paper involves the evaluation of each of these elements as well as the theory of action that underlies their association. This framework includes the following parts:

- ♦ Clarifying the **goals of the system and the theory of action** by which those goals are to be achieved
- ♦ Evaluating the **indicators** used to make accountability decisions
- ♦ Evaluating the **decision rules** (including AYP) used to determine how schools and LEAs are categorized for the purpose of distributing rewards, sanctions, and interventions
- ♦ Evaluating the **consequences** that are imposed for certain levels of performance as well as those that emerge subsequently

Given their limited resources, states may be best advised to prioritize their questions and to think of the “never-ending” validation process as a series of five-year plans. It is important for states to recognize that although validation requires the use of resources that could certainly be used elsewhere, *not* validating their systems could easily be far more expensive. The possible waste of time and resources by implementing ineffective programs, the loss of credibility, and the risk of lawsuits without a proper defense would cost much more than validation work.

States should use multiple data collection methods in their validation work. These can include, for example, extant data or data gathered via established collections, surveys, focus groups, site visits, and classroom observations

CLARIFYING GOALS AND THE THEORY OF ACTION

To clarify system goals and the theory of action is to specify what the accountability system is meant to achieve, the means by which these goals are to be reached, and the outcomes and processes that will be the focus of the validation process. This should take place as the accountability system is being designed but must also be considered as part of validation. The major questions that should be addressed in this part of the validation process include the following:

- ♦ What are the goals that this accountability system is meant to achieve?
- ♦ Who is to be held accountable for these goals in this system?
- ♦ What indicators are used to represent performance in relation to these goals?

- ♦ How and when are decisions made regarding performance toward the goals?
- ♦ What consequences are associated with different levels of performance?
- ♦ What changes are these consequences meant to affect?
- ♦ How are the intended changes thought to be related to the overall goals?

EVALUATING THE INDICATORS

Evaluation of indicators used for accountability purposes involves an investigation of their **accuracy** and meaningfulness. This investigation should address the following types of questions:

- ♦ What indicators are included in the accountability system and how is each used?
- ♦ How well do the definitions of these indicators capture what is intended?
- ♦ How reliable are the indicators that are used to make high stakes accountability decisions?

Evaluation of the assessment-based indicators encompasses the validity and reliability studies that are typically part of an assessment system. However, the use of assessment indicators for accountability purposes represents a new claim (e.g., percent proficient in reading/language arts is an indicator of school/instructional quality) and the state or LEA must gather validity evidence to support this claim.

EVALUATING THE DECISION RULES

Evaluation of the decision rules involves examining the reliability and accuracy of accountability decisions, including AYP. This examination should address the following questions:

- ♦ Do the results of the AYP model support the goals of the accountability system?
- ♦ Were the “right schools” identified for rewards, sanctions, and interventions?
- ♦ Are the results of the AYP model stable over time?

The AYP model prescribed in NCLB and most other accountability decision models are classification systems. For NCLB AYP, each school and each LEA is classified annually into the group that met AYP or the group that did not meet AYP. If the model were perfectly reliable and accurate, every school and LEA would be correctly classified. Thus, the validity of the model is based on the degree to which the AYP determinations identify the correct schools—those in greatest need of improvement—while not identifying schools that are doing an effective job educating all students.

However, no evaluation system produces perfectly accurate outcomes. Given a model that assigns schools and LEAs to two groups based on certain criteria, there are four possible outcomes for each decision:

- ♦ Identified and does need improvement

- ♦ Identified and does not need improvement
- ♦ Not identified and does need improvement
- ♦ Not identified and does not need improvement

Validation of the AYP decision involves, in part, an evaluation of the degree to which the identification status aligns with the actual need for improvement; that is, “Were the ‘right schools’ identified?” This requires the use of an external criterion, which could come from, for example, information collected systematically via surveys or site visits from a range of schools or results from non-NCLB statewide school level accountability systems.

A strategy to enhance the accuracy of AYP classifications involves the differentiation between schools by the degree to which they missed their AYP targets. For example, states could—

- ♦ distinguish between schools and LEAs that miss their AYP targets by a large margin and those that miss by very little; and/or
- ♦ distinguish between schools and LEAs that miss all or most of their AYP targets and those that miss very few or only one.

While the AYP model in NCLB allows for no such distinctions (the same sanctions must be applied to both schools), some type of scale could be developed to represent the degree to which the school failed to meet the criteria in the AYP model. This scale could be used to determine the kind or level of resources and technical assistance that a state or LEA might provide to a school.

In addition to these studies, which address accuracy of identification, states must also evaluate the reliability of their decisions, which is defined here as the stability of AYP results over time (i.e., from year to year). A strategy for evaluating stability would be to rank all the schools in the state on the criterion measure described earlier. Taking only the schools falling at the top and bottom of the rankings (e.g., the top and bottom 10% or the top and bottom 25%), determine the extent to which schools in the high need and low need groups fall in the opposite group from one year to the next. The greater the degree of such change, the lower the reliability of the model.

When conducting evaluations of accountability decision rules such as AYP, states should bear in mind the following concerns:

- ♦ The AYP determination made using the AYP model prescribed by NCLB is based on a conjunctive (non-compensatory) set of standards and use of this type of model “usually makes the accountability system much less reliable” (Gong, 2002, p. 9). In fact, it is seen as a common error in the design of effective accountability systems (Hill, 2000).
- ♦ There are many sources of error that affect the results of an accountability system. Many researchers agree that a chief contributor is sampling error (Cronbach et al., 1997; Hill, 2002; Hill & DePascale, 2002; Linn, 2001). To reduce misclassification errors that are affected by sampling error, states can employ confidence intervals with or without a minimum “n” for group size. States may also need to consider measurement error as a factor in misclassification.

EVALUATING THE CONSEQUENCES

Evaluating the consequences component of an accountability system involves consideration of how the imposed consequences are implemented and how they are related to both intended and unintended emergent changes in school and LEA functioning. It simply cannot be assumed that selecting the “right schools” and assigning them pre-specified consequences will lead to the intended reforms, will not lead to any unintended, negative consequences, and will ultimately result in achievement of the intended goals.

Studies of consequences should address the following types of questions:

- ♦ How well are rewards, sanctions, and interventions implemented?
- ♦ How do school and LEA characteristics, as well as other facets of the context, moderate the implementation of the consequences?
- ♦ To what degree are the intended actions occurring in relation to the application of rewards, sanctions, and interventions?
- ♦ To what degree are negative, unintended consequences occurring in relation to the application of rewards, sanctions, and interventions?
- ♦ To what degree are the reform activities associated with achievement of the goals of the system?

The primary concerns with regard to accountability consequences are—

- ♦ how the consequences that are imposed on schools, which include rewards, sanctions, and interventions, are implemented;
- ♦ whether imposed consequences are associated with the emergence of the intended reforms, as indicated in the state’s theory of action, and also the emergence of any negative, unintended consequences or activities; and
- ♦ whether the activities and conditions that emerge after the application of consequences are associated with the achievement of the accountability goals.
- ♦ This paper has provided a discussion of the questions surrounding validity and proposed ways of examining the validity of accountability systems based on them. A continuous exploration of findings from implementation of these and other methods will be important to refining validation methodology and continuing improvement of accountability systems.

References

- American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin*, 51, 201-238.
- American Psychological Association. (1966). *Standards for psychological tests and manuals*. Washington DC: Author.
- American Psychological Association, American Educational Research Association, & the National Council on Measurement in Education. (1974). *Standards for educational and psychological tests*. Washington, DC: American Psychological Association.
- American Psychological Association, American Educational Research Association, & the National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- American Psychological Association, American Educational Research Association, & the National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Erlbaum.
- Baker, E. L., & Linn, R. L. (2002). *Validity issues for accountability systems* [Technical Report 585]. Los Angeles: UCLA National Center for Research on Evaluation, Standards, and Student Testing.
- Baker, E. L., Linn, R. L., Herman, J. L., & Koretz, D. (2002). *Standards for educational accountability systems* [Policy Brief 5]. Los Angeles: UCLA National Center for Research on Evaluation, Standards, and Student Testing.
- Berdie, D. R., & Anderson, J. F. (1974). *Questionnaires: Design and use*. Metuchen, NJ: Scarecrow Press.
- Carlson, D. (2002). The focus of state educational accountability systems: Four methods of judging school quality and progress. In W. J. Erpenbach et al., *Incorporating multiple measures of student performance into state accountability systems—A compendium of resources* (pp. 285-297). Washington, DC: Council of Chief State School Officers.
- Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54, 297-312.
- Center on Education Policy. (2003). *From the Capitol to the classroom: State and federal efforts to implement the No Child Left Behind Act*. Washington, DC: Author.
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

- Cronbach, L. J. (1988). Five perspectives on validation argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Erlbaum.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.
- Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement*, 57, 373-399.
- Erpenbach, W. J., Forte Fast, E., & Potts, A. (2003). *Statewide educational accountability under NCLB: Central issues arising from an examination of state accountability workbooks and US Department of Education reviews under the No Child Left Behind Act of 2001*. Washington, DC: Council of Chief State School Officers.
- Garet, M., Porter, A., Desimone, L., Birman, B., & Yoon, K. (2001). What makes professional development effective? Analysis of a national sample of teachers. *American Educational Research Journal*, 38(4), 915-945.
- Garrett, H. E. (1937). *Statistics in psychology and education*. New York: Longmans, Green.
- Gong, B. (2002). *Designing school accountability systems: Towards a framework and process*. Washington, DC: Council of Chief State School Officers.
- Guion, R. M. (1977). Content validity: The source of my discontent. *Applied Psychological Measurement*, 1, 1-10.
- Guion, R. M. (1980). On Trinitarian doctrines of validity. *Professional Psychology*, 11, 385-398.
- Guilford, J. P. (1946). New standards for test evaluation. *Educational & Psychological Measurement*, 6, 427-438.
- Hanushek, E. A., & Raymond, M. E. (2002). Sorting out accountability systems. In W. Evers & H. Walberg (Eds.), *School accountability* (pp. 75-104). Palo Alto, CA: Stanford University, Hoover Press.
- Harrington-Lueker, D. (2000). When educators cheat. *School Administrator*, 57(11), 32-33, 35-39.
- Hill, R. (1997, June). *Calculating and reducing errors associated with the evaluation of adequate yearly progress*. Paper presented at the CCSSO Annual Large-Scale Assessment Conference, Colorado Springs, CO.
- Hill, R. (2000, March). *Common problems with accountability systems*. Paper presented at the Conference on Policy and Measurement Issues in Large-Scale Science and Mathematics Assessment, Washington, DC.
- Hill, R. (2001). *Issues related to the reliability of school accountability scores*. [Report on the reliability lecture from the 2000 Annual Edward F. Reidy Interactive Lecture Series]. Dover, NH: National Center for the Improvement of Educational Assessment.
- Hill, R. (2002, April). *Examining the reliability of accountability systems*. Paper presented at the Annual Conference of the American Educational Research Association, New Orleans, LA.

- Hill, R., & DePascale, C. (2002). *Determining the reliability of school scores*. Dover, NH: National Center for the Improvement of Educational Assessment.
- Hill, R. & DePascale, C. (2003). *Reliability of No Child Left Behind accountability designs*. Dover, NH: National Center for the Improvement of Educational Assessment.
- Hoffman, R. G., & Wise, L. L. (2000). *School classification accuracy final analysis plan for the Commonwealth accountability and testing system*. Alexandria, VA: HumRRO.
- Joint Committee on Standards for Educational Evaluation. (1999). *The program evaluation standards* (2nd ed.). Thousand Oaks, CA: Sage.
- Kane, T. J., & Staiger, D. O. (2002) *Volatility in school test scores: Implications for test-based accountability systems* [Brookings Papers on Education Policy]. Washington, DC: The Brookings Institution.
- Kane, T. J., Staiger, D. O., & Geppert, J. (2001). *Assessing the definition of "adequate yearly progress" in the House and Senate education bills* [unpublished manuscript].
- Lane, S., Park, C. S., & Stone, C. A. (1998). A framework for evaluating the consequence of assessment programs. *Educational Measurement: Issues and Practice*, 17(2), 24-28.
- Ligon, G. D., Jennings, J., & Clements, B. S. (2002) *Confidentiality and reliability rules for AYP: A guide for establishing rules for disaggregating and reporting assessment results*. Austin, TX: Evaluation Software Publishing.
- Linn, R. L. (2000). Assessments and accountability. *Educational Researcher*, 29(2), 4-16.
- Linn, R. L. (2001). *The design and evaluation of educational assessment and accountability systems* [CSE Technical Report 539]. Los Angeles: UCLA National Center for Research on Evaluation, Standards, and Student Testing.
- Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). *Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001*. *Educational Researcher*, 31(6), 3-16.
- Linn, R. L., Baker, E. L., & Herman, J. L. (2002, Fall). Minimum group size for measuring adequate yearly progress. *CRESST Line*, 1, 4-5.
- Linn, R. L. (1997). Evaluating the validity of assessments. *Educational Measurement: Issues and Practice*, 16(2), 14-18.
- MacQuarrie, D. (2002). The No Child Left Behind act: Regulatory guidance. *NCME Newsletter*, 10(3), 3-4.
- Marion, S. F., White, C., Carlson, D., Erpenbach, W. J., Rabinowitz, S., & Sheinker, J. (2002). *Making valid and reliable decisions in determining adequate yearly progress*. Washington, DC: Council of Chief State School Officers.

- Marion, S. F., & Gong, B. (2003, October). *Evaluating the validity of state accountability systems*. Lecture presented at the Annual Reidy Interactive Lecture Series, Nashua, NH.
- Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, 35, 1012-1027.
- Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 33-45). Hillsdale, NJ: Lawrence Erlbaum.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education and Macmillan.
- Messick, S. (1992, April). *The interplay of evidence and consequences in the validation of performance assessments*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.
- Messick, S. (1994). *Validity and psychological assessment: Validation of inferences from persons' responses and performance as scientific inquiry into score meaning*. [Educational Testing Service Research Report RR-94-45]. Princeton, NJ: Educational Testing Service.
- Palmer, S., & Coleman, A. (2003, September). *Developing a framework for policy and legal implications*. Presentation at the Council of Chief State School Officers' Workshop on Implementing AYP in States' Accountability Systems, Washington, DC.
- The Random House dictionary of the English language* [Unabridged]. (2nd ed.). (1987). New York: Random House.
- Reckase, M. (1998). Consequential validity from the test developer's perspective. *Educational Measurement: Issues and Practice*, 17(2), 13-16.
- Riddle, W. (2001). *Adequate yearly progress under the ESEA: Provisions, issues, and options regarding House and Senate versions of H.R. 1*. [CRS Report RL31035]. Washington, DC: Congressional Research Service, The Library of Congress.
- Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. New York: Academic Press.
- Shepard, L. A. (1993) Evaluating test validity. *Review of Research in Education*, 19, 405-450.
- Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational measurement: Issues and practice*, 16(2), 5-8, 13, 24.
- Smith, M.S., & O'Day, J. (1991). Systemic school reform. In S. H. Fuhrman & B. Malen (Eds.), *The politics of curriculum and testing: The 1990 yearbook of the Politics of Education Association* (pp. 233-267). Bristol, PA: Falmer Press.
- Tenopyr, M. L. (1977). Content-construct confusion. *Personnel Psychology*, 30, 47-54.

- Thum, Y. M. (2003). *No child left behind: Methodological challenges & recommendations for measuring adequate yearly progress* [CSE Technical Report 590]. Los Angeles: UCLA National Center for Research on Evaluation, Standards, and Student Testing.
- U.S. Department of Education. (2002). *Consolidated state application accountability workbook for state grants under Title IX, part C, section 9302 of the Elementary and Secondary Education Act* (Public Law 107-110). Washington, DC: U.S. Printing Office.
- U.S. Department of Education. (2003). *Response to the frequently asked question, "what if a school does not improve?"* (Retrieved November 1, 2003 from <http://www.ed.gov/nclb/accountability/schools/accountability.html>)
- Wiley, D. E. (1991). Test validity and invalidity reconsidered. In R. E. Snow & D. E. Wiley (Eds.), *Improving inquiry in the social sciences: A volume in honor of Lee J. Cronbach* (pp. 75-107). Hillsdale, NJ: Erlbaum.
- Zlatos, B. (1996). The Texas ranger of testing. *Executive Educator*, 18(2), 27-29

Appendix A: CCSSO Resources Related to Accountability System Validation

Below are resources available from CCSSO relevant to the work described in this paper. These resources are available on the CCSSO website, some for free downloading. Similar resources are available from various organizations. This list is not intended to be exhaustive of what is available.

RESOURCES ON ALIGNMENT:

Webb, N. L. (2002). *Alignment study in language arts, mathematics, science, and social studies of state standards and assessment for four states*. Washington, DC: Council of Chief State School Officers.

RESOURCES ON VALIDATION OF ACCOUNTABILITY SYSTEMS

Carlson, D. (1996). *Adequate yearly progress provisions of Title I of the Improving America's Schools Act: Issues and strategies*. Washington, DC: Council of Chief State School Officers.

Council of Chief State School Officers. (2003). *Surveys of enacted curriculum indicators CD: Surveys, reports, alignment analysis, data formats, PD guide, on-line web tool* CCSSO, Washington, DC: Author

Council of Chief State School Officers, Wisconsin Center for Education Research, & American Institutes for Research. (2003). *Surveys of enacted curriculum in English/language arts, science, and mathematics*. Washington, DC: Council of Chief State School Officers.

Erpenbach, W. J., Carlson, D., LaMarca, P. M., & Winter, P. C. (2002). *Incorporating multiple measures of student performance into state accountability systems—A compendium of resources*. Washington, DC: Council of Chief State School Officers.

Erpenbach, W. J., Forte Fast, E., & Potts, A. (2003). *Statewide educational accountability under NCLB*. Washington, DC: Council of Chief State School Officers.

Marion, S., White, C., Carlson, D., Erpenbach, W. J., Rabinowitz, S., & Sheinker, J. (2002). *Making valid and reliable decisions in determining adequate yearly progress*. Washington, DC: Council of Chief State School Officers.

Appendix B: Proposed Standards for Educational Accountability Systems

Baker, E. L., Linn, R. L., Herman, J. L., & Koretz, D. (2002). *Standards for educational accountability systems* [Policy Brief 5]. Los Angeles: UCLA National Center for Research on Evaluation, Standards, and Student Testing.

Standards on System Components

1. Accountability expectations should be made public and understandable for all participants in the system.
2. Accountability systems should employ different types of data from multiple sources.
3. Accountability systems should include data elements that allow for interpretations of student, institution, and administrative performance.
4. Accountability systems should include the performance of all students, including subgroups that historically have been difficult to assess.
5. The weighting of elements in the system, including different types of test content, and different information, should be made explicit.
6. Rules for determining adequate yearly progress of schools and individuals should be developed to avoid erroneous judgments attributable to fluctuations of the student population or errors in measurement.

Testing Standards

7. Decisions about individual students should not be made on the basis of a single test.
8. Multiple test forms should be used when there are repeated administrations of a test.
9. The validity of measures that have been administered as part of an accountability system should be documented for the various purposes of the system.
10. If tests are to help improve system performance, there should be information provided to document that test results are modifiable by quality instruction and student efforts.
11. If test data are used as a basis of rewards or sanctions, evidence of technical quality of the measures and error rates associated with misclassifications of individuals or institutions should be published.
12. Evidence of test validity for students with different language backgrounds should be made publicly available.
13. Evidence of test validity for children with disabilities should be made publicly available.
14. If tests are claimed to measure content and performance standards, analyses should document the relationship between the items and specific standards or sets of standards.

Stakes

15. Stakes for accountability systems should apply to adults and students and should be coordinated to support system goals.
16. Appeal procedures should be available to contest rewards and sanctions.
17. Stakes for results and their phase-in schedule should be made explicit at the outset of the implementation of the system.
18. Accountability systems should begin with broad, diffuse stakes and move to specific consequences for individuals and institutions as the system aligns.

Public Reporting Formats

19. System results should be made broadly available to the press, with sufficient time for reasonable analysis with clear explanations of legitimate and potential illegitimate interpretations of results.
20. Reports to LEAs and schools should promote appropriate interpretations and use of results by including multiple indicators of performance, error estimates, and performance by subgroups.

Evaluation

21. Longitudinal studies should be planned, implemented, and reporting evaluating effects of the accountability program. Minimally, questions should determine the degree to which the system—
 - a. builds capacity of staff,
 - b. affects resource allocation,
 - c. supports high-quality instruction,
 - d. promotes student equity access to education,
 - e. minimizes corruption,
 - f. affects teacher quality, recruitment, and retention, and
 - g. produces unanticipated outcomes.
22. The validity of test-based inferences should be subject to ongoing evaluation. In particular, evaluation should address—
 - a. aggregate gains in performance over time, and
 - b. impact on identifiable student and personnel groups.